Molecular Cloning of Chicken Transforming Growth Factor ß1 and Isolation of Microsatellites

by

Qiaoming Long

MSc (P.R.China)

Thesis submitted to the University of Edinburgh
for the degree of Doctor of Philosophy

Division of Molecular Biology
Roslin Institute
Roslin, Midlothian
Edinburgh EH25 9PS
United Kingdom

September, 1994

Declaration

I hereby declare that this thesis has been composed by myself, that all the work recorded in this thesis is original unless otherwise acknowledged in the text or by references and that none of the work has been submitted for another degree in the University of Edinburgh or any other university.

Qiaoming Long

Signature: Date: 6/12/96

Abstract

Transforming growth factor-ßs (TGFßs) are candidate genes in the control of chicken development and growth. It is very important to study the expression, regulation and functions of these genes.

Microsatellites are ideal DNA markers for genetic linkage studies and genome mapping. An increasing numbers of studies have demonstrated that they may also be involved in DNA homologous recombination, gene regulation and genome rearrangement in *vivo*. The $[CAG/CTG]_n$ triplets have an important role in transcriptional factors.

We have attempted to clone the 5' region of the chicken TGFß1 gene to facilitate a more detailed analysis of the expression and developmental role of TGFßs. We have developed a method for enriching microsatellites in chicken DNA libraries to facilitate the generation of polymorphic markers and the identification of potential transcriptional factors.

We have shown that chicken TGFß1 exists as an single copy gene with an upper size limit of approximately 15 to 22 kb. However, screening of a lambda phage chicken genomic library using both the TGFß1 cDNA and oligo probes failed to obtain chicken TGFß1 clones.

Procedures were developed for the construction of specific microsatellite-enriched DNA libraries. $[CA/TG]_n$-enriched genomic DNA libraries were constructed using "genetic marker selection" and DNA affinity hybridisation procedures; whereas $[CA/TG]_n$- and $[CAG/CTG]_n$-enriched liver cDNA libraries were constructed by a DNA affinity

hybridisation procedure. The frequency of positive clones in the DNA libraries constructed ranged from 0.5% to 5% depending on the type of microsatellite repeat. An enrichment of 50 fold over the classical small-insert DNA libraries has been achieved. Microsatellite-positive clones from both the genomic DNA libraries and the cDNA libraries were identified and characterised by sequencing. Microsatellite polymorphisms were studied by polymerase chain reaction and are being mapped using the EAST LANSING and COMPTON reference mapping crosses. A search of the non-redundant databases using the available information of expressed sequences revealed chicken homologues of human transcriptional factor, myocyte enhancer factor 2D (MEF2D) and human Fragile X syndrome (FMR1) as well as a substantial number of new other genes. The differential pattern of expression of the chicken MEF2D gene was examined in different chicken tissues and at various ages. A ubiquitous distribution of chicken MEF2D transcripts was revealed. Chromosome mapping and a study of the pattern of expression of these genes is underway. Comparative mapping of closely related avian species using microsatellite markers from chicken cDNA library is discussed.

The isolation of these microsatellites will facilitate the mapping of economically important quantitative traits in chicken. The mapping of expressed sequences will help to define candidate genes of these and other traits.

Acknowledgements

I would like to thank my supervisor, Dr. David W. Burt, for all his help with my research over the past four years, and the help given to me during the writing of this dissertation.

Appreciation to my university supervisor, professor Alan McNeilly, for his encouragement and continued support during my tough times.

Special thanks to Dr. Anne Ramage, Dr. Andy Law for useful discussion and making time to read my manuscript; all those who were of assistance to me, Mr. E.R. Paton, and Mr. D. Morrice.

I am very grateful to Mr Bob Scougal, Mrs Hazel Gilhooley, Dr. Helen Sang, and many others who made my stay in Edinburgh a very memorable one.

I am deeply indebted to my wife, Ying Chen, without whose love, sacrifice and patience during the last few years, it would almost be impossible for me to achieve these.

I acknowledge a scholarship from The Sino-British Technical Co-operation program.

Abbreviations

A	Adenine
ATP	Adenosine triphosphate
bp	Base pair
C	Cytosine
cAMP	Cyclic adenosine monophosphate
cM	Centi Morgan
cDNA	Complementary DNA
dH_2O	Distilled water
DMSO	Dimethylsulphoxide
DNA	Deoxyribonucleic acid
DNase	Deoxyribonuclease
DTT	Dithiothreitol
EGF	Epidermal Growth Factor
EDTA	Diaminoethanetetra-acetic disodium salt
G	Guanine
IGF	Insulin-like growth factor
kb	Kilobase pair
kDa	Kilodalton
mRNA	Messenger RNA
OD	Optical density
ODC	Ornithine Decarboxylase
PCI	Phenol/Chloroform/Iso

PCR	Polymerase chain reaction
PEG	Polyethylene glycol
QTL	Quantitative trait loci
RFLP	Restriction fragment length polymorphism
RNA	Ribonucleic acid
RNase	Ribonuclease
SDS	Sodium dodecyl sulphate
T	Thymine
TGF-ß	Transforming growth factor beta
w/v	Weight/Volume
^{32}P dCTP	Deoxycytidine 5'-[α-^{32}P] triphosphate, triethyl-ammonium salt
^{32}P ATP	Adenosine 5'-[γ-^{32}P] triphosphate, triethyl-ammonium salt
^{35}S dATP	Deoxyadenosine 5'-[α^{35}S] thiotriphosphate, tri-ethylammonium salt

Contents

Chapter 1

Introduction

1.1 Chicken Transforming Growth Factor-ß

Many genes are expressed coordinately and the proteins coded by these genes interact with one another in the development of chicken muscle and skeleton. Genes that may be involved include growth factors, hormones and their receptors, and transcriptional factors. Transforming growth factor-ßs (TGFß) have been shown playing a crucial role in the cellular proliferation and differentiation during chondrogenesis, myogenesis, osteogenesis, hematopoiesis and adipogenesis (Jakowlew and Lechleider *et al.*, 1992). They are appropriate "candidate genes" for controlling growth rate in chicken. We, therefore, are interested in studying the gene structure, expression and regulation, and functions of TGFßs.

1.1.1 Main Functions of TGFß in vivo

The transforming growth factor beta peptides are a group of functionally distinct and structurally related regulatory proteins. Results of *in situ* hybridisation and immunohistochemical staining have shown ubiquitous expression of TGFß in both embryo and adult tissues (Yang *et al.*, 1990). Various biological functions have been inferred from these patterns relating to the control of cell proliferation, cell adhesion and cell phenotype (Massagué, 1990). These are as summarised in Table 1.1.

1.1.2 TGFß Genes and Peptides of Mammals

So far, in mammals three types of TGFß peptides have been identified, named TGFß1, TGFß2 and TGFß3 (Derynck *et al.*, 1987ab; Derynck *et al.*, 1988; Kondaiah *et al.*, 1988). A novel type of TGFß (TGFß5) was identified in *Xenopus laevis* (Kondaiah *et al.*, 1990). However, it is not clear, as yet, whether this represents a new

gene or an ancestral gene related to TGFß1/TGFß3 (Burt and Paton, 1992). All these peptides share a common precursor structure: a hydrophobic N-terminal signal sequence, a pro-region and a C-terminal bioactive domain (see Figure 3.1). As far as gene structure is concerned, all members of the family have a highly conserved exon-intron structure: seven exons interrupted by six introns.

Table 1.1 Summary of the Major Functions of TGFßs

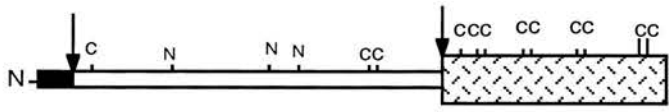
Type of Function	**Affected Cells/Tissues/Components**
Cell proliferation (+/-)	Epithelial, endothelial, fibroblast, neuronal, lymphoid and hematopoietic cells
Cell Adhesion (+/-) (cell-cell interactions)	Up regulates: extracellular matrix proteins, their receptors and protease inhibitors; down regulates: proteolytic enzymes
Cell phenotype (+/-) (differentiation)	Adipocyte, myoblast, chondroblast, osteoblast, epithelial, endothelial, hematopoietic cells, steriodogenic cells.

(+/-): positive or negative effects.

1.1.3 TGFß Genes in Chicken

However in chicken, four distinct TGFßs (TGFß1-4) have been reported (Jakowlew *et al.*, 1988abc). In 1988, Jakowlew *et al.* apparently cloned the chicken homologue of human TGFß1. Soon after publishing this chicken TGFß1 cDNA sequence, experimental evidence both in and out of our lab (unpublished data) brought this conclusion in doubt. First, the TGFß1 cDNA sequence showed 100% homology with porcine TGFß1. For two species that diverged about 300 million years ago (Nei, 1987), this would seem to be highly unusual. Second, we used an oligo probe designed using the published chicken TGFß1 cDNA sequence to screen a chicken

genomic library but failed to obtain any positive clones. Furthermore, TGFß2, 3 and 4 mRNAs, but not TGFß1 mRNA were detected in cultured chondrocytes and myocytes (Jakowlew *et al.*, 1990), and adipocytes (Burt *et al.*, 1992) both by northern blotting and RT-PCR.

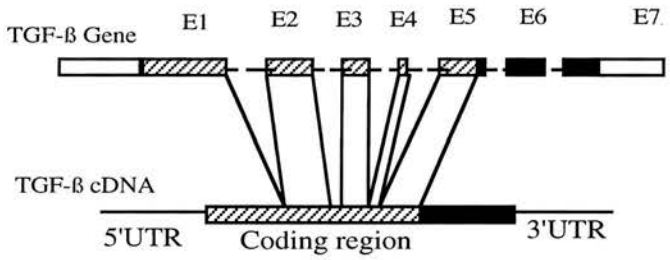


(A) TGFß precursor structure

N glycosylation sites

C conserved cysteines

↓ cleavage sites



(B) Intron-exon structure of TGFß gene

Figure 1.1 TGFß Gene Family. (A) Schematic representation of TGFß precursor; (B) Intron-exon structure of TGFß genes.

The chicken TGFß2 and TGFß3 sequences share the highly conserved precursor structure with their mammalian homologues. But TGFß4 which has only been identified in chicken has certain unique features. Unlike the previously described TGFßs which are 390 to 414 amino acids long, the predicted precursor protein of TGFß4 is only 304 amino acids, and does not appear to contain a signal peptide sequence. In addition, it is the only TGFß to have an insertion of two amino acids near

the N-terminus of the processed peptide which would results in a 114 amino acid mature protein (Jakowlew *et al.*, 1988a). Burt and Jakowlew re-examined the nucleotide sequence of the chicken TGFß4 cDNA using the Taquenase system rather than the Sequenase system originally used, and revealed six errors in the original publication (Burt and Jakowlew, 1992). The revised cDNA sequence has a longer reading frame and encodes a typical TGFß precursor. In contrast to the previously published conclusions, it does encode a signal peptide and therefore, has the potential for being a secreted peptide.

A review of these data suggested that the characterised chicken TGFß4 may be the chicken homologue of mammalian TGFß1. In order to test this hypothesis, Burt and Paton (1992) calculated a molecular phylogeny of the transforming growth factor beta gene family based on a comparison of cDNA sequences. The proposed phylogenetic tree shows that the family evolved from a common ancestral gene. This model suggests that the present TGFß gene family consists of four members: TGFß1, TGFß2, TGFß3 and TGFß5. The southern hybridisation data in this study also suggests that the proteins for mammalian TGFß1 and chicken TGFß4 are the products of homologous rather than duplicated genes. These studies provided strong evidence to support the hypothesis that there exist three types of TGFß genes or peptides in chicken. It is now generally accepted that chicken TGFß4 is the homologous gene of mammalian TGFß1. Throughout this chapter, this will be assumed and TGFß1 will be used.

Based on these arguments, Burt (unpublished data) suggested that Jakowlew *et al.* investigate the sequence of the published chicken TGFß1 cDNA. Enquiries revealed that this sequence was in fact a porcine TGFß1 cDNA (Kondaiah *et al*, 1988), possibly the result of a clone mixed-up.

1.1.4 Differential expression of TGFßs and Characterisation of TGFß Genes Promoters

A large number of studies show that expression of the different isoforms of TGFß vary from cell to cell. They are regulated differentially both in embryogenesis and in adult tissues and also by a variety of growth factors and hormones including epidermal growth factor, retinoic acid, dexamethasone, tamoxifen, phorbol esters and the TGFßs themselves (Jakowlew and Lechleider *et al.*, 1992; Glick *et al.*, 1991; Wakefield *et al.*, 1990). To investigate this further, Roberts *et al.* (1991) cloned and characterised the promoters for human TGFß1, 2 and 3 genes. Significant difference have been found between these promoters, for example, whereas the TGFß1 promoter has no TATAA box and is regulated principally by AP-1 sites, both TGFß2 and TGFß3 promoters have TATAA boxes, AP-2 sites, and cAMP responsive elements. It has been demonstrated that the basis for the differential expression of the human TGFßs lies, in part, in the differences of the 5' flanking sequences of these genes.

In an attempt to study the differential expression and regulation of TGFß genes in chicken, the promoter regions for chicken TGFß2 and TGFß3 have been cloned and characterised (Burt *et al.*, 1991; Jakowlew and Cubert *et al.*, 1992). Whilst the chicken TGFß2 and TGFß3 promoters show little sequence homology to each other, they are structurally similar; like the human promoters, both contain TATA box, CRE and AP-2 sequence motifs near the transcriptional initiation site. Sequence comparisons between the chicken and human TGFß2 and TGFß3 promoters revealed conserved sequences 111 bp conserved region in the TGFß2 promoters. An 86 bp sequence surrounding the transcriptional start and a 156 bp sequence in the 5' untranslated region are conserved in the TGFß3 promoter.

1.1.5 Methodology of TGFß Studies

There are a number of ways to perform studies on the functions of TGFßs. (1) Gene cloning and promoter characterisation. This will facilitate further studies on the expression and regulation of TGFßs through *in vitro* cell culture experiment or whole animal experiment. However, the results from these studies are not direct evidence of a role for TGFßs. (2) Gene knocking-out and transgenic studies. This will provide direct evidence of the involvement of TGFßs in the *in vivo* biological process investigated. However, it is technically difficult to generate transgenic chicken. (3) Genetic linkage studies. To study the possible involvement of TGFßs in the development and growth control in chicken, segregation of TGFß genes with microsatellite markers used to map quantitative trait loci (QTL) can be followed in a broiler-layer reference cross. This will lead to the proposal of "candidate genes", possibly including TGFßs.

1.2 Molecular Biology of Microsatellites

1.2.1 Classification and Distribution of Tandem Repetitive DNA Sequences

The existence of repetitive sequences in eucaryotic genomes was discovered two decades ago by DNA reassociation experiments (Britten and Kohne, 1968). Eucaryotic repetitive sequences are classified into three types:

Satellite DNA is composed of very long blocks of highly repeated DNA, in which simple sequences are repeated thousands of times. In many cases, these repeating sequences have compositions unlike that of most of the genome and so are easily separated by centrifuging slightly fragmented DNA in a caesium chloride density gradient (Estoup *et al.*, 1993; Watson *et al.*, 1987). Satellite DNAs are usually

localised near the centromere or telomeres of a chromosome. The sequence is characteristically long and the inter-repeat variability is high.

Minisatellites refer to sequences which are composed of tandem repeats of 9 to 64 bp motifs and have a total length ranging from 0.1 to 7 kb (Estoup *et al.*, 1993; Jeffreys *et al.*, 1985). They are also frequently located near the centromeric and telomeric regions. Compared with satellite DNA, minisatellites are relatively short in length and have a lower inter-repeat variability.

Microsatellites correspond to tandemly repeated motifs of 1 to 5 bp spanning a total length usually less than 0.2 kb (Estoup *et al.*, 1993; Rassmann *et al.*, 1981; Tautz and Renz, 1984). Inter-repeat variability is almost zero. In contrast with satellite and minisatellite DNA, *in situ* hybridisation and genomic southern blotting show that microsatellites are randomly distributed throughout the genome, but underrepresented near the centromeric or telomeric regions (Wong *et al.*, 1990; Hamada and Kakunaga, 1982; Hamada *et al.*, 1982).

Of the three types of repetitive sequences described above, microsatellites have attracted most attention recently because of their potential as highly polymorphic genetic markers for genome mapping, their possible involvement in homologous recombination and regulation of gene expression.

1.2.2 Evolution of Microsatellite Sequences

Base substitution, amplification, unequal sister chromatid exchange and natural selection are the four processes that modify a string of nucleotides in sequence and length (Wolfgang, 1989). In order to account for various structural features of tandemly repeated DNA sequences, two mechanisms have been proposed for the generation of tandem repetitive sequences: unequal crossing over (UCO) and slippage

strand mispairing (SSM) (Smith, 1976; Levinson and Gutman, 1987). UCO and SSM are two different mechanisms which either work alone or together. They have some features in common. For example, they both can generate duplications and deletions in DNA in a manner dependent on homologous base pairing. On the other hand, the two mechanisms differ in that SSM is an intrahelical event, involving the two strands of a single DNA duplex, whereas UCO is an interhelical event involving DNA molecules from two different chromosomes or sister chromatids. This places special constraints on UCO, since it can only take place during chromosome alignment in cell division and will be dependent on such factors as the rate of chiasma formation. SSM on the other hand, ought to be free of such constraints and could potentially occur whenever unpaired loops form, during DNA repair as well as replication. SSM might therefore be expected to be an inherently more frequent event (Levinson and Gutman, 1987).

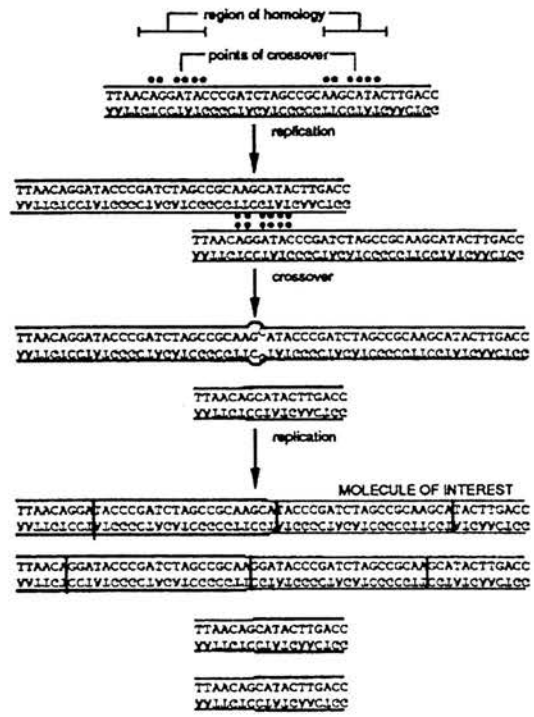


Figure 1.2 Unequal sister chromatid cross over (UCO). Heavy and light lines indicate DNA strands which derive from the upper and lower strand of the starting molecule, respectively; the points of crossover are indicated in the starting sequence; the duplicated sequences are delineated by vertical bars in the final molecules. (Smith, 1976).

(1) Unequal Crossing Over (UCO)

Smith (1976) proposed that DNA whose sequence was not maintained by natural selection ("neutral DNA") would exhibit tandem repetitive patterns as a result of frequent unequal crossing over (UCO), if strand exchange in the recombination process was assumed dependent on sequence similarity. Figure 1.2 shows the generation of repetitive sequence through sister chromatid unequal cross-over. Wolfgang and Cho (1994) compared the data on microsatellite, minisatellite and satellite DNA with the results from their simulation study. It was found that the properties of minisatellite and satellite DNAs resembled the simulated structures very closely. This suggests that unequal crossing-over is indeed a dominant long-range ordering force for the formation of these two types of repetitive sequences. However, they found that the microsatellite data did not fit the sequence structures from simulation using the proposed model, suggesting that UCO does not act on these short tandem arrays, at least not as the dominant force.

(2) Slipped Strand Mispairing(SSM)

The structural characteristics of microsatellites are a simple repeat unit (1-5 bps) and low inter-repeat variability. Such arrays appear to be too short for UCO which involves two separate molecules. What then might be the mechanism by which microsatellites were generated and evolved? To answer this question, Levinson and Gutman (1987) analysed different sources of naturally occurring simple repetitive sequences, and proposed that slipped-strand mispairing (SSM) is the major mechanism for the formation of microsatellite sequences.

SSM, as shown in Figure 1.3, involves local denaturation and displacement of the strands of a DNA duplex followed by mispairing of complementary bases at the site

of an existing short tandem repeat. The simplest consequences of this mispairing, when followed by replication or repair, can lead to insertions or deletions of one or several of the short repeat units. The consequences can thus provide a coherent explanation for the origin and evolution of simple repetitive sequences in genomic DNA. As Levinson and Gutman argued , SSM is much more likely to be the major factor in the initial expansion of short repeated motifs. After initial expansion, simple tandem repeats may be predisposed to further expansion by UCO or other inter-strand events.

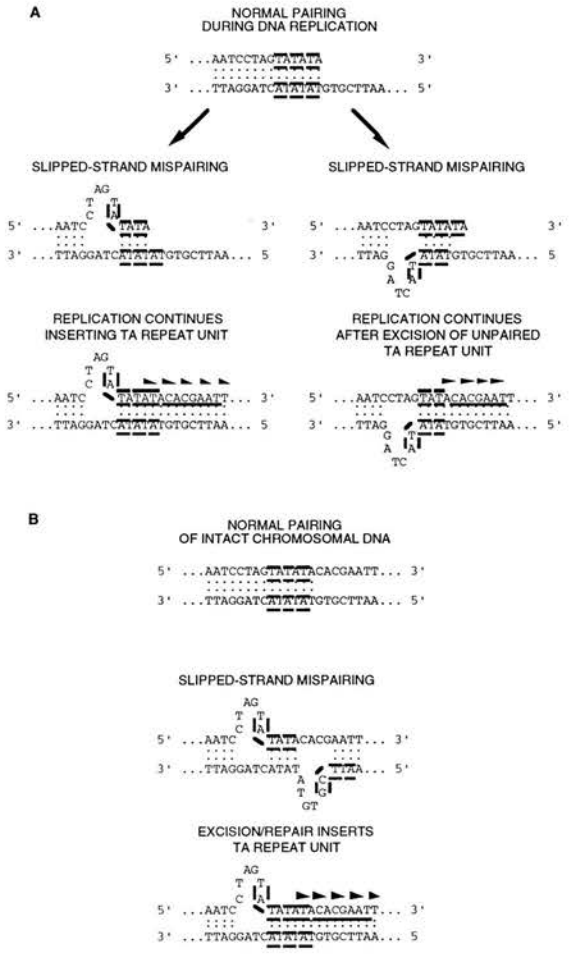


Figure 1.3 Slipped-strand mispairing (SSM). (A) Slipped-strand mispairing during DNA replication; (B) Slipped-strand mispairing of intact chromosomal DNA.

Fornage and colleagues (1992) examined the allele frequency distribution of a $(TG)_n(AG)_m$ microsatellite in the Apolipoprotein C-11 gene. Valdes *et al.* (1993) summarised available data on the allele frequencies at 108 microsatellite loci in human populations. They compared the observed frequency distribution pattern to those generated by a simulation of the so called "Stepwise Mutation" model, and found that both were consistent with the results generated from the model under the assumption that mutations cause an increase or decrease in repeat number by one. For example, there was always one allele with intermediate size and highest distribution frequency, although the distribution was quite irregular, showing bimodal, trimodal or even quadrimodal distribution. According to the theory that the age of an allele was proportional to its frequency (Fornage *et al.*, 1992), this allele was likely to be the ancestor. From this ancestor, the other alleles were generated through increasing or decreasing one or several repeat units. The progressive and continuous decrease of the frequencies from the oldest allele strongly supports the hypothesis that alleles were derived from one another by SSM, a mechanism creating mutation of one single repeat unit by deletion or insertion. However, from Fornage's examination of the frequency distribution at $(TG)_n(AG)_m$ loci, one can also see that UCO was acting to some extent, since each of the three modes comprised one frequent allele and one very rare allele adjacent in size. No alleles of intermediate size were found between the first three modes.

(3) Other Possible Mechanisms

Roger (1983) proposed that some of the $[CA/TG]_n$ repeats in the genome might be generated through a totally different mechanism. He observed that many of the repeat blocks were flanked by terminal repeats of 6 to 8 base pairs. For example, CCTAACTA TT(CT)29(CA)30(GA)4G CCTAACTA in the mouse immunoglobulin CδH–Cδ3 intron. Terminal repeats were taken as possible evidence of insertion at a

staggered break, in the manner of transposons. Therefore he proposed that some of the $[CA/TG]_n$ repeats were inserted into certain sites of the genome. The most promising source of these inserted $[CA/TG]_n$ fragments seems to be at the ends of chromosomes.

1.2.3 Possible Involvement of Microsatellites in In Vivo Molecular Process

$[CA/TG]_n$ and $[CG/GC]_n$ microsatellites have been shown to adopt a novel DNA conformation, possibly left-handed Z DNA conformation, in negatively supercoiled plasmids (Hamada and Kakunaga, 1982; Lancillotti *et al.*, 1987; Nordheim and Rich, 1983ab; Kladde *et al.*, 1994). In addition, using these DNAs on a chromotograph matrix, various types of DNA binding proteins have been isolated (Fishel *et al.*, 1988). Recently, microsatellites have been found to be the target binding sites for a number of nuclear proteins (Fishel *et al.*, 1993). These studies have raised several possibilities concerning the functions of microsatellites *in vivo*.

(1) Microsatellites and Homologous Recombination

Microsatellites could participate in or promote homologous recombination by two possible mechanisms.

First, various types of microsatellites are evenly dispersed throughout the genome of eucaryotic organisms. Such a distribution of a large number of simple sequence tandem repeats would simply increase the likelihood that a pairing partner would encounter a homologous sequence. Homologous recombination would be facilitated by the binding of a specific recombination enzyme. The recent electron microscope observation by Gaillard and Strauss (1994) that $[CA/TG]_n$ sequences form a tetra-stranded structure bound by nucleoprotein HMG1 and HMG2, supports this hypothesis.

Second, microsatellite sequences are potential Z-DNA forming sequences (Hamada and Kakunaga, 1982; Nordheim and Rich, 1983ab; KLysik *et al.*, 1981). The formation of Z-DNA, on the one hand, would create a single-stranded region at the B-Z border area by DNA unwinding; and on the other, facilitate the binding of specific Z-DNA proteins. Therefore microsatellites are potential target sites for recombinatory enzymes, such as Rec1, in the genome. Recently, a human enzyme with strand transferase activity was partially purified from the extracts of T lymphoblast by Z-DNA affinity chromotography (Fishel *et al.*, 1988). The ability to isolate such a protein strongly implies that Z-DNA is involved in genetic recombination in mammalian cells.

Evidence to date suggests that microsatellites can promote homologous recombination. Most evidence comes from the examination of *in vivo* DNA exchange, gene duplication and other molecular events. For example, $[CA/TG]_n$ microsatellites have been shown to be involved in gene duplication events at the human α-globin (Proufoot and Maniatis, 1980), β-globin (Treco *et al.*, 1985; Treco and Arnheim, 1986), immunoglobulin-IgA C_H (Flanagan *et al.*, 1984) and human fetal globin loci (Slightom *et al.*, 1980). A [CAGG] tetra-nucleotide repeat was associated with the high recombination frequency within the Eß gene at the MHC locus (Streinmetz *et al.*, 1986). The most convincing evidence comes from the fusion-function assays with simian virus 40 (SV40). Both synthetic and genomic $[CA/TG]_n$ repetitive sequences have been introduced into SV40. The results of these studies demonstrated that microsatellites may promote homologous recombination (Stringer, 1985; Bullock *et al.*, 1986; Bernues *et al.*, 1991; Murphy and Stringer, 1986 Wahls *et al.*, 1990). Based on these results, microsatellite sites in the genome are recognised as potential 'hot spots' of genetic recombination.

The dispersion of large numbers of microsatellites in eucaryotic genomes and the possibility that microsatellites can promote homologous recombination suggest they have a role to play in genome evolution. Gross and Garrard (1986) revealed that $[CA/TG]_n$ microsatellites were not detectable in the genomes of Eubacteria, Archaebacteria and Mitochondria. These results suggest that such repetitive sequences may play a role in packaging and DNA condensation in eucaryotic chromosome (Stallings *et al.*, 1991). Considering the high abundance of microsatellites in eucaryotic genomes and their possible involvement in homologous recombination, it seems likely that they may play a role in eucaryotic genome rearrangements.

(2) Microsatellites and Transcriptional Regulation of Genes

Evidence supporting the involvement of microsatellites in transcriptional regulation comes from sequence examination of transcriptional enhancers. Microsatellites are present within the regulatory region of simian virus (SV40) (Ondek *et al.*, 1988). Now microsatellite sequences have been found in the regulatory regions of a number of cloned genes, for example, the downstream of the joining region (JH-Cμ) in mouse immunoglobulin heavy chain genes (Gillies *et al.*, 1983; Banerji *et al.*, 1983); the 5' flanking region of the human metallothionein-IIA gene (Karin *et al.*, 1984) and the upstream region of the rat α-tropomysin gene (Herrera *et al.*, 1990).

Hamada *et al.* (1984) introduced both synthetic and genomic poly[CA/TG] tracts into a plasmid vector and investigated the effect of these microsatellites on transcription control by monitoring CAT activity. It was found that a poly[CA/TG] element increased CAT gene expression about 10 times over controls. Very recently, Wang *et al.* (1994) showed that the $[CTG]_n$ triplet repeat in the myotonic dystrophy gene preferentially facilitated nucleosome formation when expanded to a certain threshold.

These results support the idea that microsatellites may be involved in transcriptional regulation.

1.3 Microsatellites and Genome Mapping

1.3.1 Why Do We Need a Gene Map ?

There are several major driving forces behind the development of a gene map, a linear display of gene order.

(1) Map-based positional cloning of genes: Higher eucaryotic genomes contain tens of thousands of genes, for example, the human genome has been estimated to have 60,000 to 70,000 genes (Weber, 1994). Without knowledge of its chromosome location, a gene can only be identified and cloned based on its biochemical activity ("*function-based cloning*"). For instance, expression cloning was used in cloning a rat gene that stimulates dibasic and neutral amino acid transport (Wells and Hediger, 1992). However, detailed knowledge of the number and arrangement of genes in the genome will facilitate "*map-based cloning*" which does not depend on any knowledge of function at the biochemical level. Map-based cloning of a gene is achieved by locating molecular markers closely linked to the locus.

(2) Comparative gene mapping: A comparison of the maps across different species will provide knowledge on chromosome evolution and facilitate the search for candidate genes affecting a biological phenotype. For example, the common ancestor of all mammals appeared about 80 million years ago (Ralph, 1991). During this period, the arrangement and number of chromosomes have changed by processes such as chromosome translocations, insertions/deletions and fusions. However, the

arrangement of genes on some chromosome are often identical among different species. This is more common for species that share a recent ancestor, e.g. cattle and sheep. Historically, gene mapping in human and mouse has been more detailed and systematic than in other species. Comparing gene maps of the poorly-mapped species with these map-rich species can lead to the proposal of candidate genes within an homologous chromosome region, which may fit a mapped biological trait.

(3) Genetic analysis of polygenic or quantitative traits: Phenotypic variation of a large number of physiological and economic traits, such as milk production, egg production and growth rate, are determined by a relatively large number of genes known as "quantitative trait loci" or QTL. The observed phenotype is also influenced by a wide variety of environmental factors. These make a classical Mendelian genetic analysis of such traits very difficult. This has led to the development of the "biometrical" approach to study QTL. However, biometrical analysis does not provide direct information as to the number of QTL, their individual effects, their dominance and epistatic relationships, or their pleiotripic effects on other traits. The inability to directly follow the inheritance of individual QTL limits the rate of genetic improvement by selection, particularly when dealing with traits showing low heritability, sex-limited or late developmental expression, or that are difficult to evaluate in live animals. From specific line crosses, marker-trait linkages can be detected, and genes or chromosome regions affecting quantitative traits can be identified (Anderson *et al.*, 1994).

1.3.2 Types of Maps and Mapping Methods

Two types of genetic maps have been assembled in both human and livestock species: (1) physical and (2) genetic linkage or recombination map. A physical map is based on the physical assignment of genes onto individual chromosomes using cytogenetic

methods or by ordering large-insert clones (e.g. yeast artificial chromosome, YAC). A genetic linkage map is based on the frequency of meiotic exchange among polymorphic markers in sexual crosses (O'Brien, 1993; Tuggle *et al.*, 1994). The methods used to develop genetic maps are shown in Table 1.2.

Table 1.2 Types of genetic maps and mapping tools

Types of Map	**Method Used to Develop the Map**
Physical Map	Somatic cell hybridisation
	In situ hybridisation
	Ordering YACs
Linkage Map	Restriction fragment length polymorphism analysis (RFLP)
	DNA fingerprinting analysis using minisatellite (VNTR)
	Microsatellite polymorphism analysis (MS)
	Random amplified polymorphic DNA analysis (RAPD)
	Single-stranded conformation polymorphism analysis (SSCP)

For technical reasons, gene assignment in the chicken depends far less on somatic cell and *in situ* hybridisation, and far more on meiotic mapping approaches than gene assignment in human and other mammals. (1) The complex nature of the chicken karyotype (38 pairs of autosomes, Z and W sex-chromosomes in each diploid cell) and the fact that the chicken genome has a number of microchromosomes (30 pairs) make the chromosomes difficult to distinguish by classical cytogenetic techniques (Levin *et al.*, 1994). (2) Somatic cell hybrids are not widely available in chicken. Nevertheless, 35 genes have been assigned to chicken macrochromosomes and 20 genes to microchromosomes by *in situ* hybridisation (Burt *et al.*, personal communication). (3) The design and analysis of a chicken backcross is also a simple

task and large numbers of backcross progeny are available for genetic linkage analysis.

1.3.3 DNA markers

To construct a linkage map, polymorphic markers are needed. For the past decades, linkage mapping in poultry has relied on physiological markers, such as feather colour, comb shape, isozymes and immunological markers (Levin *et al.*, 1994). Recently, advances in DNA technology has led to the development of DNA-based markers.

(1) Advantages of molecular or DNA-based markers

DNA markers offer several advantages over physiological markers (Tuggle *et al.*, 1994):

1) DNA markers based on DNA variation are usually codominant meaning that each allele can be detected. This allows the unambiguous determination of genotypes, since both homozygous classes and heterozygous animals can be distinguished.
2) DNA marker technology is both rapid and reliable.
3) There is high genetic variation at the DNA level, which can be detected with a number of techniques.
4) DNA tests can be performed very early in the life of the animal, even at the embryonic stage. Thus genetic testing and subsequent selection decisions can be taken before significant investment is made in an individual animal, even for quantitative traits which are expressed late in life or in the future progeny of the animal.

(2) Comparison of various types of genetic markers

Several types of DNA markers have been developed. A comparison of their advantages and disadvantages is given in Table 1.3.

Table 1.3 Comparison of different types of genetic markers used in genome mapping

Marker	Advantages	Disadvantages
RFLP	-Does not require sequence information -Large gene regions can be analysed	-Slow (filter hybridisation) -Uses large amounts of DNA -Labour and materials costs high
VNTR	-Large number of loci can be investigated at once -Useful in investigating parentage identification	-Dominant marker only, so some genetic information can be lost -Require hybridisation
MS	-Rapid, no hybridisations -Most polymorphic of all currently used markers -Large numbers and even distribution in genome -Amenable for automation	-Require DNA sequence information -Variability not likely to be biologically relevant -High cost
RAPD	-Rapid, no hybridisations -Low cost analysis -Simple laboratory analysis	-Dominant marker only, some information can be lost -Sensitive to PCR conditions, reproducibility low
SSCP	-Rapid, no hybridisations -Analysis is very sensitive	-Requires gene sequence information

RFLP: restriction fragment length polymorphism; VNTR: variable number of tandem repeats; MS: microsatellite; RAPD: random amplification polymorphism of DNA; SSCP: single stranded conformation polymorphism.

1.3.4 Status of Linkage Mapping in Farm Animals

Construction of a genetic map in animals began in 1913 (Sturtevent, 1913). In fact, the chicken was the first vertebrate species to be studied (Hutt and Lamoreux, 1940). However the idea of mapping complex genomes (human and livestock) using DNA-

based markers started in the mid 1980s. Now genetic maps exist for more than 30 species (O'Brien, 1993). A summary of the current status is given in Table 1.4.

Table 1.4 Current status of genetic linkage maps of human and major domestic animals

Species	Haploid Number	Mapped Loci	Linkage Groups
Cattle	30	450	29
Sheep	27	140	21
Pig	19	600	19
Chicken	38	461	32*, 36**
Mouse	20	2800	20
Human	23	6552	24

* In the East Lansing linkage map; ** in the Compton linkage map.

The chicken has a number of advantages for construction of a linkage map (Bumstead and Palyga, 1992): (1) the genome is small relative to other domestic species (see Table 1.2); (2) the production of progeny is high; (3) it is possible to prepare good quality DNA in large quantity from small volumes of blood, since the red blood cells are nucleated.

The earliest chicken linkage map was constructed by Hutt and Lamoreux (Hutt and Lamoreux, 1940). In 1992, a preliminary linkage map of the chicken was produced consisting of 100 RFLP mapped on 18 linkage groups (Bumstead and Palyga, 1992). During the last two years, there has been dramatic progress in producing the chicken genetic linkage map. Two internationally recognised mapping crosses, the East Lansing mapping cross (Crittendem *et al.*, 1992) and the Compton mapping cross (Bumstead and Palyga, 1992) have been set up. Several laboratories are developing

microsatellite markers for chicken genome mapping. A review of the genetic markers in the current chicken linkage map is given in Table 1.5.

Table 1.5 Linkage markers and linkage groups of the current chicken genetic linkage map (Burt, unpublished data)

Marker Type	**East Lansing Cross**	**Compton Cross**	**Both**	**Total**
Classical	10	2	2	10
CR1	47	0	0	47
EST	5	24	4	25
EVL	7	31	0	38
Genes	37	35	10	62
Genes/Microsatellites	37	35	10	62
Microsatellites	39	26	15	50
Minisatellites	33	25	16	42
Random Clones	22	100	18	104
RAPD	68	0	0	68
Total	**282**	**247**	**68**	**461**
Linkage Group	32	36		50

Notes: CR1, chicken repeat 1; EST, expressed sequence tag; EVL, Endogenous viral loci; RAPD, random amplified polymorphic DNA.

Microsatellites have been extensively used in the development of dense genetic maps in human, mouse, porcine, cattle and other domestic farm animals. A summary of the microsatellite-based genetic linkage maps of human and major domestic animals is given in Table 1.6.

Table 1.6 Progress of linkage maps using microsatellites (Weissenbach *et al.*, 1992; Rohrer *et al.*, 1994; Bishop *et al.*, 1994; Dietrich *et al.*, 1992).

Species	Microsatellite Loci	Haploid Numbers	Linkage Groups
Human	814	23	23
Pig	376	19	24
Cattle	292	30	30
Chicken	50	38	31
Mouse	317	20	20

An international collaboration is being coordinated to develop a genetic linkage map of the chicken based on microsatellites and other DNA markers. The aim is to use microsatellite markers to scan the entire genome for genes that control quantitative traits (QTLs). This will demand a large number of microsatellite markers.

1.4 Aims of the Thesis

The aims of the work described in this thesis are as follows:

1) Molecular cloning of chicken TGFß1 gene to facilitate the study of TGFß1 gene expression regulation
2) Developing a method for enriching microsatellite clones in DNA libraries. This will facilitate the efficient isolation and characterisation of large numbers of microsatellites from the chicken genome.
3) Isolating and characterising microsatellite clones from the genomic and cDNA libraries. The large numbers of microsatellite markers obtained will facilitate whole genome scanning and QTL mapping.

Chapter 2

Materials and Methods

2.1 Materials

2.1.1 Chicken Genomic Library

Two chicken genomic libraries of different resources were used in the isolation of chicken transforming growth factor beta 1 (TGFß1) genomic clones. Library I was constructed and maintained in our laboratory by cloning size-selected *Sau3A* partial digested fragments into the *BamHI* site of phage lambda EMBL3. The average size of the inserts was 15 kb. Library II was purchased from Strategene. This was constructed by cloning *Sau3A* partial digested 10 day old male White Leghorn DNA into the *XhoI* site of Lambda Fix II vector. The insert size was 9-22 kb.

2.1.2 Bacterial Host Strains

Strain	Source	Genotype
NM621	Working Laboratory	recD1009 thy+P1 transductant of SB204 (White et al., 1988)
C600	Working Laboratory	e^{14}(mcrA),supE44,thi-1,thr-1,leuB6,laeY1,tonA21 (Appleyard, 1954)
CJ 236	Bio-Rad	dut-1,ung-1,thi-1,relA1,pcJ105(cm^r)
NM522	Working Laboratory	(hsd-mcrB)Δ5supE(lac-pro)Δ/F'pro+lacI0 lacZ M15 (Gough and Murray, 1983)
DH5a	Working Laboratory	F-φ80dlacZM15D (lacZYA-argF)U169end A1recl
XL1-Blue	Strategene	hsd17(r_K-m_K+)deoR thi-1 supE44λ-gyrA96re1A1 recA1,endA1,gyrA96,thi-1,hsdR17,supE44

2.1.3 Antibiotics

All anti-biotics were dissolved in distilled water (unless indicated), sterilised by filtration, and stored in aliquots at -20^{0}C.

Anti-biotics	Stock Concentration.	Working Concentration	Notes
Ampicillin	50 mg/ml	50 µg/ml	
Chloramphenicol	30 mg/ml	15-30 µg/ml	in 100% ethanol
Kanamycin	50 mg/ml	70 µg/ml	
Tetracillin	15 mg/ml	12 µg/ml	

2.1.4 Chicken TGF-β1 Specific cDNA Probes

Chicken TGFß1 cDNA (in pUC19) was kindly provided by Professor J.M.Connor (Department of Medical Genetics, University of Glasgow). The cDNA was double-digested with *BamHI* and *PflMI*. Restricted fragments were separated on 1% 'LMP' agarose gel. All restriction fragments were the anticipated size. cDNA probes were then prepared by fragment preparation (M&M 2.2.12): Probe 1: 230 bp, PflMI-BamHI, nucleotide 49-277; Probe 2: 690 bp, BamHI-BamHI, nucleotide 277-965; Probe 3: 918 bp, PflMI-BamHI, nucleotide 49-965.

2.2 Methods

2.2.1 Generation of DNA probes

Radiolabelling of Synthetic Oligonucleotide: Synthetic oligonucleotides were phosphorylated at the 5' terminus by transfer of ^{32}P from [$\gamma^{32}P$]ATP using standard methods. 50 pmol of synthetic oligonucleotides was mixed with 5 x kinase buffer, 10 µl; liquid [$\gamma^{32}P$]ATP (specific activity=5,000 Ci/mmol), 5-10 µl at the concentration of 10 mCi/ml; T_4 polynucleotide kinase, 20 units. dH_2O was added to 50 µl. The labelling reaction was carried out by incubating at 37^0C for 45 to 60 minutes. The

reaction was stopped by adding 0.5 M EDTA, 2 µl followed by one round of PCI extraction and ethanol precipitation with linear polyacrylamide as a DNA carrier. The pellet was washed once with 70% ethanol, dried under vacuum at room temperature and resuspended in 200 µl of TE buffer. Then 2 µl of the labelled probe was spotted on a Whatman filter, dried down at 65^0C for 5 to 10 minutes, the specific activity was counted in a scintillation counter using a scintillation vial containing the filter and 5 ml of scintillation buffer.

Radiolabelling of Double-stranded DNA Fragment: 50-100 ng of DNA fragment was mixed with 32 µl dH_2O, boiled for 5-7 minutes to denature, and immediately chilled on ice. Then the following reagents were added in the stated order: OLB buffer, 10 µl; bovine serum albumen (10 mg/ml), 2 µl; [α^{32}P] dCTP, 5 µl; Klenow Fragment, 5 units. The mixture was incubated at room temperature for 2-4 hours or overnight. The reaction mostly took place in the first 3-12 hours. The reaction was then stopped by mixing with 2 µl of 0.5 M EDTA and 48 µl of dH_2O. The labelled probe was recovered by ethanol precipitation and subsequently washed with 70% ethanol. After drying down under vacuum at room temperature, the pellet was resuspended in 200 µl TE buffer, from which 2 µl was removed and spotted on a Whatman filter for counting and calculation of the specific activity. The probe was boiled for 7-10 minutes to denature before adding to the hybridisation mixture.

2.2.2 Isolation of Genomic Clones

Preparation of plating cells (Sambrook, Fritsch and Maniatis, 1989): A single bacterial colony was placed into 100 ml L. Broth containing 10 mM $MgSO_4$ and incubated at 37^0C with moderate shaking for 2 to 3 hours until the OD_{550} of the culture reached 0.5 (about 2.5 x 10^8 cells/ml). Cells were pelleted and resuspended in

50 ml of 10 mM $MgSO_4$ at 37^0C for 45 minutes with moderate agitation. The plating cell suspension was stored at 4^0C and used within 2 weeks.

Screening of genomic DNA library (Sambrook, Fritsch and Maniatis, 1989): For 10^6 phages, we distributed 20,000 phages per plate and a total of 50 plates. Recombinant phages were infected into 200 µl NM621 competent cells at 37^0C for 20 minutes and plated onto a L-agar plate (90 mm size). These plates were incubated at 37^0C for 6-8 hours until the plaques just make contact with one another. Plates were then chilled at 4^0C O/N to harden the top agarose. A replica of the plaques was obtained by gently laying a nitrocellulose filter onto the surface of the Top agarose for 30-60 seconds. The filters were peeled off and dried at room temperature for 1 hour. Phage particles were denatured by immersing the filters (DNA side up) into denaturing solution (0.5 M NaOH, 1.5 M NaCI) for 30-60 seconds, followed by immersing into neutralising solution (1.5 M NaCI, 0.5 M Tris-HCI, [pH 7.4]) for 1 minute and finally rinsing in 2 x SSC. The DNA was irreversibly bound to the filter by baking at 80^0C under vacuum for 2 hours. The filters containing immobilised DNA were prehybridised overnight at 37^0C in prehybridisation buffer (see section 2.2.6) containing poly A, poly C, denatured SS DNA and λDNA at final concentrations of 10 µg, 10 µg, 100 µg and 1 µg per ml respectively. The poly A, poly C, SS DNA and λDNA were boiled for 5 minutes before adding to the prehybridisation buffer. Phage DNAs were then hybridised with a ^{32}P labelled probe (10^6 cpm/ml) in hybridisation buffer at 37^0C for 48 hours. After hybridisation the filters were rinsed in 2 x SSC twice at room temperature and then washed twice in 2 x SSC, 0.1% SDS for 30 minutes each at room temperature, followed by two times in 0.5 x SSC, 0.1% SDS at 55^0C for 15 minutes each. Filters were dried at room temperature and autoradiographed O/N at -70^0C. Positive plaques were picked using a glass tube and resuspended in 1 ml of phage buffer containing a drop of chloroform.

2.2.3 Phage Stock Preparation

Plate lysate stock: A healthy pure positive plaque was picked and resuspended in 1 ml of phage buffer containing 1 drop of chloroform, which yields a titre of $10^6\phi$/ml. 0.1 ml of the suspension (10^5 pfu) was mixed with 0.2 ml of freshly prepared C600 cells; incubate in 37^{0}C-water bath 20 minutes for adsorption. Then 3 ml of BBL top agarose was added and the total mixture was poured onto a moist fresh L-agar/agarose plate. The infected cells were grown up for 6-8 hours at 37^{0}C until the plaques touched one another. The phages were harvested in 3 ml phage buffer and shaking at 4^{0}C for several hours. 0.1 ml chloroform was added and mixed with the lysate. Cells and debris were removed from the plate lysate by centrifugation.

Liquid lysate stock: 10 ml overnight-culture of C600 cells was combined with 200 ml L-broth containing 10 mM $MgCl_2$ in a 2-litre flask and incubated at 37^{0}C with vigorous shaking (180 rpm) for 1-2 hours until OD_{600} reached 0.5 ($2x10^8$ cell / ml). Phages from the plate lysate was added at M.O.I = 1.0 (4×10^{10} phages), and incubation was continued for 3 hours. 0.5 ml of chloroform was added and the culture was shaked for another 10 minutes. The liquid lysate was harvested by pelleting the cells/debris and titred for phage concentration.

2.2.4 Phage DNA Preparation

Rapid mini-preparation: 1.5 ml of phage stock from liquid lysate was centrifuged at 13,000 g for 5 minutes to remove bacterial debris, and treated with RNase A (final concentration = 50 μg/ml) for 30 minutes at 37^{0}C. 1.2 ml of the supernatant was then centrifuged at 75 K in a Beckman TL100 ultra-centrifuge for 30 minutes. The pellet was dissolved by vortex in 100 μl GuHCl solution (6 M GuHCl, 1% sarkosyl, 25 mM EDTA pH 8.0, 0.5 M ammonium acetate). Proteinase K was then added to final concentration of 1 μg/μl and the solution was incubated for 2 hours at 50^{0}C. The

phage DNA was precipitated with one volume of isopropanol, washed with 70% ethanol and dissolved in 30 µl of TE. The resuspension was further treated with 120 units of pre-boiled (20 minutes) RNase T1, which made it essentially free from RNA. Ammonium acetate was added to a final concentration of 2.5 M and the RNA-free DNA was isopropanol precipitated as above, and resuspended in 25 µl TE.

Large scale preparation: DNase I/RNase I was added to a final concentration of 1 µg/ml to the phage stock obtained from liquid lysate and incubated for 30 minutes at room temperature. NaCl was added at 40 g/litre and the mixture was ice-bathed for 1 hour. Cell debris was then pelleted and removed by centrifuging at 10 K for 15 minutes at 4^0C. To pellet the phage particles, PEG-8000 (polyethylene glycol, MW 8000) was added to 10% (W/V); The mixture was left at 4^0C in stationary overnight or longer, followed by centrifuging at 5 K for 15 minutes at 4^0C. Meanwhile CsCl centrifuge block gradient was set in a ultra-clear centrifuge tube. Following centrifuging, the pellet was dissolved in 1 ml of phage buffer which was then loaded on top of the CsCl block gradient. The DNA was extracted from the ultra-centrifuge tube using a needle. The remaining CsCl salts, RNAs and proteins were removed through dialysis at 4^0C in TE buffer and treatment with RNase A (final concentration=10 µg/ml), and Pronase (final concentration=1 µg/ml) respectively. After dialysis in 0.1 M NaCl/1 mM EDTA and 0.002% triton X-100 at 37^0C overnight, phage DNA was extracted twice with two volumes of phenol, and finally dialysed against TE buffer at 4^0C overnight. DNA concentration and relative purity was determined by measuring OD_{260} and OD_{280}. The standard formula used for measuring DNA concentration was µg of DNA=50 x dilution x OD_{260}.

2.2.5 Polymerase Chain Reaction

Reactions were carried out in a total volume of 50 µl containing 50 mM KCl, 10 mM Tris-HCl (pH 8.3), 1.5 mM $MgCl_2$, 0.01% (w/v) gelatin, 0.1% (v/v) Triton X-100, 200 µM of each dNTP, 50 pmol of each primer, and 0.5 units of Taq DNA polymerase (Promega). Evaporation within the tube was prevented by the addition of 30 µl of mineral oil overlay. The reaction mixture was subjected to 25-30 cycles of 1 min at 94^0C, 2 mins at 45^0-60^0C, and 3 mins at 72^0C followed by a final extension of 10 mins at 72^0C.

2.2.6 DNA Blotting and Probing

Vacuum blotting: DNA was digested (0.5-1.0 µg/digest) with a series of restriction endonucleases, electrophoresed in 1.0% agarose gel. After electrophoresis, the gels were treated twice, 15 minutes each time in the order of fractionating buffer (0.25 M HCl), denaturing buffer (0.5 M NaOH/1.5 M NaCl) and neutralising buffer (1 M NH_4Ac/0.02 M NaOH). The final round of neutralising buffer was used as transferring buffer. The depurinated and denatured DNAs were transferred onto Zetabind membrane (Bio-Rad) under vacuum (40-50 cm.H_2O) for about 30 to 60 minutes. The Zetabind membrane was then treated for 10 minutes on 0.4 M NaOH-soaked filter pad and rinsed in 2 x SSC before air drying. DNA was irreversibly bound to the membrane by baking at 80^0C for two hours. Prehybridisation was performed for 15-30 minutes. Then high specific activity [^{32}P]-labelled probe was added, hybridisation was performed for 4-12 hours depending on probe. Following hybridisation, the membrane was rinsed in 2 x SSC at room temperature; washed in 2 x SSC/0.1% SDS twice, 30 minutes each at 55^0C; and finally washed in 0.5-1 x SSC/0.1% SDS once at 55^0C for 15 minutes. Filters were dried at room temperature and autoradiographed.

Capillary blotting: DNA was digested with restriction endonucleases and electrophoresed in 5% acrylamide gel (0.8 mm). After electrophoresis, the acrylamide gel was treated for 30 minutes with denaturing buffer and neutralising buffer respectively, and subsequently released from the gel plate onto a filter paper. The gel was cut into appropriate size and placed on top of the sponge which was soaked fully in neutralising buffer. The gel was covered with Zetabind membrane, on which bunch of tissue paper was placed. The denatured DNA in the acrylamide gel would be transferred onto the Zetabind membrane by capillary reaction by leaving overnight with certain amount of pressure evenly applied on the top of the gel. After blotting, the membrane was treated, prehybridisation and hybridisation were performed as described previously in vacuum blotting.

2.2.7 Subcloning

Plasmid or M13 DNA was cleaved with restriction enzymes. The reaction mixture was PCI, CI extracted, precipitated in dH_2O and treated with Calf Intestine Phosphorylase (CIP). Meanwhile, foreign DNA was digested with corresponding enzymes, PCI and CI extracted and precipitated in dH_2O. The cleaved vector and foreign DNA were ligated together by incubating at 4^0-16^0C for 8 to 12 hours. After ligation, the reaction mixture was heated at 65^0C for 5 minutes to denature the enzyme, followed by PCI and CI extraction.

2.2.8 Transformation

Heat shock protocol: 10 ml of SOB medium was inoculated with a single *E. coli* colony from a freshly restreaked L-agar plate and incubated at 37^0C with shaking over night. The small culture was transfered into a flask containing 100 ml of L-broth, and cultured at 37^0C with moderate agitation until cell density reached 4-7 x 10^7 viable

cells/ml (OD_{600}=0.5). The cells were pelleted by centrifugation at 2000-3000 rpm for 15 minutes at 4^0C. The cells were resuspended by moderate vortexing in a volume of RF1 that was 1/3 of the volume collected. The cells were incubated on ice for 15 minutes to 2 hours and pelleted as before. The pelleted cells were resuspend in RF2 to 1/12.5 of the original volume, and incubated on ice for 15 minutes. 100 µl of the cell suspension was aliquoted to separate tube. DNA solution in a volume of 2-10 µl was added and mixed. The DNA/cell mixture was incubated on ice for 1 hour and subsequently heat shocked by placing the tube in a 42^0C water bath for 90 seconds and immediately returned to ice for 20 minutes. 400 µl of SOC medium was added and incubated at 37^0C with moderate agitation for 30 minutes to 1 hour. 100 µl was spreaded on X-gal-Ampicillin-L-agar plate. The plates were incubated overnight at 37^0C.

Electroporation: *E. coli* cells were grown as described previously until OD_{600} reached 0.5-0.8. The culture was chilled on ice for 15 minutes, then centrifuged at 2000-3000 rpm for 15 minutes at 4^0C. The cells were washed for three times in sterilised and ice-chilled 10% glycerol solution, and finally resuspended in 150-200 µl of the same solution. 60 µl of the cell suspension was removed to a microfuge tube, DNA was added in a volume of 5 µl. The DNA and cells were mixed well by pippeting and left on ice for 1 minute. The mixture was then transferred into an electroporate cuvette . Immediately after electroporation, 1 ml of SOC medium which has been equilibrated to room temperature was added. The mixture was transfered into a plastic tube and incubated at 37^0C with vigorous shaking for 1 hour. Finally, 10-20 µl was spreaded on X-gal-ampicillin-agar plate.

2.2.9 Plasmid DNA preparation

Mini-preparation: (Qiawell-8 Plasmid DNA Purification Kit)**:** 5 ml of L-broth containing ampicillin (50 µg/ml) was inoculated with a single colony, and grown over night at 37^0C with vigorous shaking. The cells were harvested by spinning at 2000-3000 rpm for 15 minutes and resuspended in 300 µl of P1 buffer. 300 µl of P2 buffer was added and mixed gently by inverting the tube 4-6 times by hand to avoid shearing of chromosome DNA, and incubated at room temperature for 5 minutes. 300 µl of chilled buffer P3 was added and mixed gently as in the previous step, and incubated on ice for 10 minutes. Cell debris and *E. coli* chromosome DNA was then removed by centrifuging for 15 minutes at maximum speed (13,000 rpm). The recovered supernatant was applied to a Qiawell-8 strip and allowed to enter the membrane disk. The Qiawell-8 strip was washed three times with 1 ml of buffer QC. The DNA was eluted out with 0.5 ml of buffer QF and precipitated by adding 0.7 volume of isopropanol and immediately spinning for 30 minutes at room temperature.

Plasmid Midi-Maxi preparation: This protocol is designed for 30-150 ml (midi-prep) or 150-500 ml (Maxi-prep) culture of *E. coli* cells. After centrifugation, the cell pellet was resuspended in 4 or 10 ml of buffer P1. A volume of 4 or 10 ml of buffer P2 was added and mixed gently and incubated at room temperature for 5 minutes. Then 4 or 10 ml of buffer P3 was added and mixed immediately but gently and centrifuge at full speed for 30 minutes at 4^0C. Remove supernatant promptly, and centrifuged again for 10 minutes to obtain a particle-free lysate. A Qiagen tip-100 or 500 was equilibrated with 3 or 10 ml of buffer QBT. The supernatant was applied to the Qiagen tip and allowed to enter the resin by gravity. The Qiagen tip-100 or 500 was washed with 2 x 5 ml or 3 x 10 ml of buffer QC. DNA was eluted with 5 or 15 ml of buffer QF, precipitated and washed as described previously.

2.2.10 DNA Sequencing and Sequence Analysis

Single-stranded DNA Sequencing: To generate single-stranded DNA template, DNA of interest was cloned into M13 vector and transfected into a male *E. coli* host, NM522. Recombinant plaques were picked using a sterilised toothpick to infect 2 ml of diluted NM522 cells which was incubated at 37^0C for 7-8 hours. Single-stranded DNA was prepared by denaturing phages with and subsequent removal of CATB. Both Sequenase and Taq polymerase systems were used in the sequencing reaction. For each template, 1 µl of primer, 2 µl of reaction buffer and 10 µl of DNA (approximately 1 µg or 0.5 pmol of DNA) in dH_2O or 1/10 diluted TE were combined. The mixture was heated to 70^0C for 2 minutes, then allowed to cool slowly to room temperature in 70^0C-water bath, over a period of 30 minutes. After annealing was complete, the tube was placed on ice. To the annealed primer-template, 2 µl of labelling mixture, 0.5 µl of [α^{35}S]dATP, 2 µl of diluted Taq DNA polymerase were mixed thoroughly and incubated for 5 minutes at 45^0C. Meanwhile, 4 µl of each termination mix was placed into 4 tubes labelled G, A, T and C. As soon as the labelling reaction was completed, 4 µl of the labelling mix was aliquoted to each of the four termination mixes and incubated in a 70^0C water bath for 5 minutes. The sequencing reaction was stopped by adding 4 µl of stop solution. Just before loading on a sequencing gel, the samples were heated to 70^0C for 5 minutes to denature DNA.

Double-stranded DNA Sequencing: About 3-5 µg of plasmid DNA was linearized by digestion with appropriate endonuclease in a volume of 20 µl. 4 µl was checked on a mini-gel and the remaining was PCI extracted. The DNA pellet was resuspended in 17.6 µl of 1/10 diluted TE. 4.4 µl of sequenase reaction buffer was added to a total volume of 22 µl. The double-stranded DNA was converted into single-stranded DNA by incubating with 0.5 µl T_7 gene 6 exonuclease for 30 minutes at 37^0C. The enzyme was inactivated by heating for 10 minutes at 70^0C. 4 µl was checked on a mini-gel.

Then 9 µl was removed to 1 µl of primer. The primer/DNA mixture was heated to 65^0C for 2 minutes, cooled down to room temperature in 65^0C-water bath and placed on ice after annealing was completed. 1µl of DTT, 2 µl of diluted labelling mix, 0.5 µl of [α^{35}S]dATP and 2 µl of diluted sequenase were added and mixed thoroughly. The mixture was incubated for 2-5 minutes at room temperature. 2.5 µl each termination mix was placed into 4 tubes labelled T, C, G, and A. As soon as the labelling reaction is completed, 3.5 µl of the labelling mix was aliquoted into each of the four labelled tube, spun down to mix and incubated at 37^0C for 3-5 minutes. The reactions were terminated by adding 4 µl of stop solution.

2.2.11 Preparation of Genomic DNA from Blood

100 µl of chicken blood was mixed with 300 µl of reagent A (10 mM Tris, 320 mM Sucrose, 5 mM $MgCl_2$, 1% Triton X-100, adjusted to pH 8.0) for 5 minutes at room temperature. The mixture was then centrifuged at 1300 g for 5 minutes. The pellet was gently resuspend in 1 ml of reagent B (400 mM Tris, pH 8.0; 60 mM EDTA; 150 mM NaCl; 1% SDS, added after autoclaving). Then 1 µl RNase solution was added and incubated at 37^0C for 30 minutes. 250 µl of sodium perchlorate was added and shaken at 37^0C for 15 minutes. The tube was incubated in a 65^0C water bath for 25 minutes shaking occasionally. 300 µl of chloroform which had been stored at -20^0C was added and mixed well by inversion for 2 minutes. The sample was centrifuged at full speed in a microfuge for 5 minutes. Then 800 µl of the upper, DNA bearing, phase was removed to a fresh tube. 0.8 volume of isopropanol was added to precipitate the DNA. The DNA was recovered by spinning for 5 minutes. The pellet was washed with 1 ml 70% ethanol, dried briefly and resuspended in 500 µl of TE.

2.2.12 DNA Fragment Preparation

DNA was digested to completion with appropriate restriction endonuclease and PCI extracted. Fragments of different sizes were separated by running on 1% low-melting agarose gel. Gel slices were cut and weighed in a 15 ml tube. QB buffer was added to a volume of 5 ml/g gel. The tube was placed in a 65^0C water bath for 10 minutes to melt gel. The solution was transferred to a new tube containing 2 g/g gel of urea, mixed by inverting the tube several times and placed in 37^0C water bath until urea was dissolved. Meanwhile, a Qiagen tip-5 column was equilibrated with 1 ml QBT. The sample was run onto the column. The column was washed with 6 x 1ml QB buffer. DNA was eluted out with 800 µl of QF buffer, precipitated and washed as described previously.

2.2.13 Extraction of mRNA from Tissue

ß-Mercaptoethanol, 41 µl and GTC extraction buffer [4 M guanidine thiocyanate/25 mM sodium citrate (pH7.1)], 1 ml were added into a 15 ml sterile tube. The tube containing the buffer was weighed. Working as quickly as possible, chicken liver was extracted and placed into a mortar and pestle. Liquid nitrogen was added and the liver tissue was homogenised by grinding until no visible fragments remained. The homogenised tissue was transferred into the tube containing extraction buffer, and mixed thoroughly by inversion. The tube containing the tissue in extraction buffer was weighed. The tissue mass was calculated by subtracting the weight obtained previously from the new weight. Referring to the standard chart, the amount of biotinylated oligo dT necessary for the tissue mass was determined. 2 ml of the pre-heated dilution buffer was removed to a sterile tube and mixed with 41 µl of ß-mercaptoethanol. The mixed solution was added to the homogenate and mixed thoroughly by inversion. Appropriate amouts of biotinylated oligo dT determined was

added and mixed well by shaking. The mixture was incubated at 70^0C for five minutes. The homogenate was then transferred to a clean sterile tube and centrifuged at 12,000 g for ten minutes at room temperature. During centrifugation, the SA-PMPs were resuspended by gently rocking the bottle. Appropriate amounts of pre-washed particles was added to the homogenate and mixed by inversion. The homogenate/SA-PMP mixture was incubated at room temperature for two minutes. The SA-PMPs were captured using the magnetic stand in the horizontal position until the homogenate clears. The particles were resuspended in 1 ml 0.5 x SSC and transfered to a clean tube. The washing was repeated twice. After the final wash, appropriate volume of nuclease-free water was added and the particles were magnetically captured. The mRNA was pelleted by adding 0.1 volume of 3 M sodium acetate and 1 volume of isopropanol and incubating at -20^0C overnight followed by centrifuging. The pellet was washed with 70% ethanol. The concentration and purity of the mRNA were determined by measuring OD_{260} and OD_{280}.

2.2.14 cDNA Synthesis

To a sterile RNase-free tube, oligo $d[T]_{25}$ primers and the mRNA sample (use 0.5 µg primer/µg mRNA) was added in a total volume of 15 µl. The mixture was heated to 70^0C for five minutes and cooled down to room temperature. Then the following components was added: 5 x first strand buffer, RNase inhibitor, 40 mM sodium pyrophosphate, AMV reverse transcriptase and dH_2O. The reaction was carried out at 42^0C for one hour. After incubation, the tube was placed on ice. The following components were added to the first strand reaction mixture: 10 x second strand buffer, *E.coli* DNA polymerase I, *E.coli* RNase H and dH_2O to the optimal volume. The mixture was incubated at 14^0C for two hours. 0.1 volume of 200 mM EDTA, was added and the mixture was heated at 70^0C for ten minutes. Then the reaction mixture was extracted by PCI and CI, and precipitated in TE or water.

2.2.15 Genetic Marker Selection

Size-selected DNA was ligated into *SmaI* digested and dephosphorylated pTZ phagemid vector. The ligation mixture was treated with restriction enzyme *SmaI*. Then recombinant DNA was transformed into *E. coli* CJ236 cell by electroporation. About 10^5 colonies were plated out onto ten plates (150 x 15 mm) containing ampicillin (100 μg/ml; LB-Amp) and incubated at 37^0C overnight. 10 mls of 2 x YT with ampicillin (100 μg/ml) was added to the surface of each plate. The plates were rotated gently at 37^0C for two hours, after which the resulting cell suspensions were collected. The plates were rinsed with 10 mls of 2 x YT and the aliquots were combined. The cells were pelleted and washed once with 10 ml 2 x YT. Then 10 ml of LB-Amp was inoculated with 50 to 100 μl of the above cell mix and grown to saturation at 37^0C. Then 50 ml of 2 x YT with ampicillin (100 μg/ml) and chloramphenicol (15 μg/ml) was inoculated with 0.5 to 1 ml of the saturated cell culture and incubated at 37^0C until OD_{600} about 0.3, i.e. 1 x 10^7 CFU/ml. Helper phages, M13K07 were added to M.O.I=20 (i.e. 20 phages/cell). Incubation was continued at 37^0C for one hour. Then kanamycin was added to a final concentration of 70 μg/ml. Incubation was continued at 37^0C overnight. Cell debris and bacterial chromosome DNA was removed by spinning the culture at 12 K for five minutes. The supernatant was transferred to a fresh tube and recentrifuged. At this point, the phagemid titre was checked by transfecting CJ236 cells. The second supernatant was transferred to a fresh tube. The phagemid solution was incubated with 150 μg RNase at room temperature for 30 minutes. Then 1/4 volume 3.5 M ammonium acetate/20% PEG-8000 was added, mixed well and incubated on ice for 30 minutes. Phagemids were collected by centrifugation at 12K rpm for 15 minutes and resuspended in 200 μl high salt buffer (300 mM NaCl, 100 mM Tris, pH8.0 and 1 mM EDTA), chilled on ice for 30 minutes. Insolubles were removed by centrifuging for 2 minutes. The supernatant was transferred to a fresh tube. The entire 200 μl phagemid stock was

extracted twice with equal volume of neutralised phenol, once with PCI, several times with CI until no visible interface was seen. The aqueous phases were pooled together, and 0.1 volume of 7.8 M ammonium acetate and 2.5 volume of ethanol were added. The mixture was placed at -70^{0}C for more than thirty minutes followed by centrifuging at 4^{0}C for 15 minutes. The pellet was washed with 90% ethanol. The pellete was resuspended in 20 µl TE, and treated with Exonuclease III (Promega).

Primer extension reactions were carried out by combining 0.1-0.3 µg of DNA, 1-2 pmols of 5'-phosphorylated $[CA]_{15}$ or $[CCT]_{10}$ primers, 10 x Taq buffer (Promega), 1 µl; and 200 µM each dNTP in a final volume of 100 µl. Samples were heated to 94^{0}C for 10 minutes and then cooled to 75^{0}C. Five units of Taq polymerase (Promega) were then added to each sample and the mixture was incubated for an additional 30 minutes at 75^{0}C. The reactions were terminated by extraction with one volume of PCI, followed by extraction with two volumes of CI and ethanol precipitated with 20 µg of glycogen, 0.3 M sodium acetate and 3 volumes of ethanol. The precipitates were resuspended in 9 µl of dH_2O, 1 µl of 10 x ligase buffer and incubated with T_4 DNA ligase at 37^{0}C for 1 hour. The ligation mixture was PCI and CI extracted and ethanol precipitated in 10 µl dH_2O. The ligation mixture was transformed by electroporation into XL1-Blue cells. These transformants were referred to as the genetic marker-selected library.

2.2.16 DNA Affinity Hybridization and Capture

Preparation of Lone Linker: 1.5 nmoles of each of the two oligonucleotides (5' GAGATATTAGAATTCTACTC 3'; 5' GAGTAGAATTCTAATAT 3') was mixed in a volume of 38 µl. The mixed oligonucleotides were heated at 90^{0}C for two minutes and quickly placed on ice. Then 10 µl of 5 x kinase buffer, 20 units of

polynucleotide kinase was added and the mixture was incubated at 37^0C for one hour. The enzyme was inactivated by heating to 70^0C for five minutes, followed by slowly cooling to room temperature in water bath.

Ligation of Lone Linker: The lone linkers were ligated to size-selected DNAs by incubating at 18^0C overnight in a total volume of 20 µl containing 5 pmoles of DNA, 30 pmoles of lone-linkers, 2 µl of 10 x Ligase buffer and 200 units of T_4 DNA ligase. After the ligation reaction, the enzyme was inactivated by heating to 70^0C for five minutes. The mixture was PCI and CI extracted, and precipitated in an appropriate volume of water giving concentration about 1 to 2 ng/µl.

Subtractive-Hybridisation: The linker-attached DNA was amplified by PCR in a total volume of 100 µl containing DNA, 1-2 ng; 200 µM each dNTPs; 10 x Taq buffer, 10 µl; Taq, 2.5 units and primers (5' GAGATATTAGAATTCTACTC 3'), 100 pmoles. The cycling conditions used were 1 minute at 94^0C, 2 minutes at 55^0C, 3 minutes at 72^0C, 30 cycles followed by 10 minutes at 72^0C. Excess PCR primers were later removed from the reaction mixture by passing through a Microcon-100 filter (Promega). The DNA concentration was checked by measuring OD_{260} and running samples on a gel.

The DNA affinity hybridisation was carried out by combining 2 µg of DNA, 2 µl of 5 M NaCl; 1 µl of 10% SDS; 1 µl of 1 M Tris-HCl pH8.0; 0.5 µl of 0.2 M EDTA, and dH_2O to 100 µl. The mixture was heated to 95^0C for five minutes, followed by adding 375 ng biotinylated primers $[CA]_{15}$ or $[CAG]_{10}$, and continued at 95^0C for one minute. Then the mixture was transferred to 70^0C for ten minutes and placed immediately on ice. Pre-washed streptavidin coated magnetic beads (Promega) were added in a volume of 30 µl, and the total volume was adjusted to 200 µl by adding the following together: 5 M NaCl, 4 µl; 10% SDS, 2 µl; 1 M Tris-HCl (pH8.0), 2 µl; 0.2 M EDTA, 1 µl and dH_2O, 60 µl. The mixture was incubated at room temperature for

one hour on a rotary mixer and washed five times with 200 µl wash buffer A [0.1 M NaCl/0.1% SDS/10 mM Tris-HCl (pH8.0)/1 mM EDTA(pH8.0)] at room temperature, once at 70^{0}C. Washed twice with buffer B [0.1 M NaCl/10 mM Tris-HCl(pH8.0)/1 mM EDTA(pH8.0)] at 70^{0}C. Bound DNA was eluted with three rounds of 100 µl buffer C (0.1 M NaCl/0.1 N NaOH). The eluted DNA was precipitated and resuspended in 10 µl dH_2O.

The selected DNA was PCR amplified as described previously. The PCR products were treated at 37^{0}C for 30 minutes, 68^{0}C for 10 minutes in a reaction containing 5 mM EDTA(pH8.0), 1 mM Tris-HCl, 0.5% SDS and 5-10 µg of Proteinase K. After PCI and CI extraction, the precipitated DNA was digested overnight with restriction enzyme *EcoRI* to create ends for cloning. The EcoRI digested DNA was then cloned into pBluescript (+). The ligation mixture was transformed into XL1-Blue cells. The transformants were referred to as the DNA affinity hybridisation selected library.

2.2.17 Bacterial Colony Hybridisation

About 150 to 300 recombinant colonies were plated onto L-Agar-ampicillin plates and incubated overnight at 37^{0}C. Plates were air-dried for 30 minutes at room temperature. A replica was made by placing a High-Bond membrane on top of each of the plates for 30 to 60 seconds and peeled off gently. The replica was put on 2 x SSC/5% SDS buffer-soaked filter paper for 1 to 2 minutes, then bacterial colonies were denatured by microwaving at full scale for two minutes. The membrane was prehybridised at 37^0-48^{0}C for 30 minutes in prehybridisation buffer (20% Formamide/0.25 M NaH_2PO_4(pH7.2)/0.25 M NaCI/1% SDS/1 mM EDTA) and hybridised 4-5 hours in the same buffer but with ^{32}P-labelled probes.

2.2.18 Cycle Sequencing Using 373 Automated DNA Sequencer

The sequencing reaction pre-mixture was made by adding the following components in the order shown: 5 x TACs buffer, 4 µl; dNTPs, 1 µl; each of the DyeDeoxy™ terminator, 1 µl and Taq polymerase (ABI), 0.5 µl (2.5 units). The sequencing reaction mixture was prepared by adding the following components in a 0.5 ml microcentrifuge tube: pre-mixture, 9.5 µl; primer (3.2 pmol/µl), 1 µl; DNA (0.25 µg/µl), 4 µl; and dH_2O, 5 µl. The cycle conditions used were as follows: 94^0C, 30 seconds; 55^0C, 1 minutes; 64^0C, 4 minutes; 25 cycles. After sequencing reaction, the reaction mixture was purified using the QIAquick-spin DTR Kit for dye terminator removal. The DNA was precipitated with 20 µl of 3 M NaAC (pH5.1) and 550 µl 96% ethanol, and resuspended in 5 µl of loading buffer (1 µl 50 mM EDTA in 5 µl formamide). Before loading, samples were heated at 90^0C for two minutes to denature DNA.

Chapter 3

Molecular Cloning of Chicken Transforming Growth Factor ß1

3.1 Introduction

Based on the structural conservation of the TGFß2 and TGFß3 promoters among different species, and the differences between the human TGFß2/TGFß3 promoters and the human TGFß1 promoter, we would predict that the chicken TGFß1 promoter would be similarly unique. The cloning and characterisation of the 5' region of the chicken TGFß1 gene will (1) facilitate the study of TGFß1 gene regulation; (2) allow a comparison of the mammalian and chicken TGFß1 promoters, which would highlight conserved regions that might be important for TGFß1 promoter function. Therefore we screened chicken genomic libraries to isolate genomic clones that contain the 5' region of chicken TGFß1. We found that although the chicken TGFß1 gene exists as a single copy gene in the chicken genome, it is, however, not represented in the chicken genomic libraries used. Possible causes of this phenomenon are discussed.

3.2 Genomic Blot Analysis of Chicken TGFß1 Gene

Chicken genomic DNA (10 µg) was digested with the restriction enzymes *BamHI, BglI, EcoRI, HindIII, KpnI, SalI, SmaI, SstI* and *XhoI*. Restriction fragments were separated on a 1% agarose gel and blotted onto Hybond membrane. The blot was probed with a [α^{32}P]dCTP labeled chicken TGFß1 full length cDNA (1400 bp). As shown in Figure 3.1, a single fragment with the approximate size of 15-22 kb was seen in the *XhoI*-digested genomic DNA. This suggests that the TGFß1 gene exists as a single copy gene in the chicken genome with an upper size limit of approximately 15 to 22 kb. The size of the exon-spanning region of this gene is therefore very similar to that of TGFß3 (15.6 kb) (Dey, 1991). Comparison of the restriction map of the TGFß3 gene and the crude restriction map of the TGFß1 drawn from the genomic

blotting (see Figure 3.2) has ruled out the possibility of TGFß1-TGFß3 cross-hybridisation.

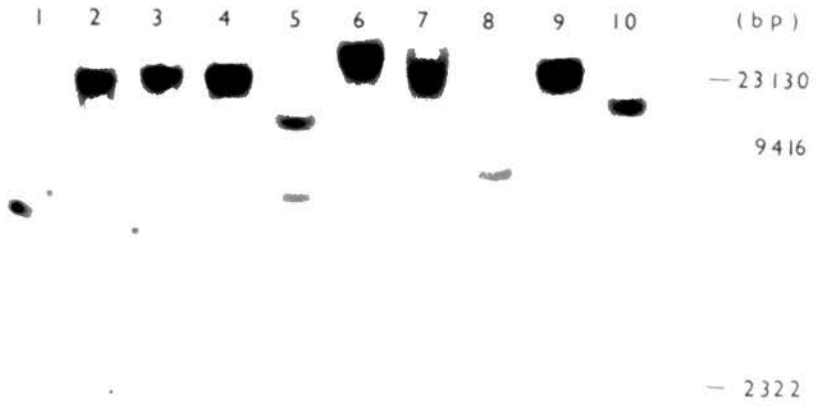


Figure 3.1 Genomic southern blot analysis of chicken TGFß1 gene. The blot was probed with [α^{32}p] dCTP TGFß1 cDNA. Hybridisation and subsequent washing were carried out as described in M&M 2.2.6. From left to right: *BamHI, BglI, EcoRI, HindIII, KpnI, SalI, SmaI, SstI, XhoI.*

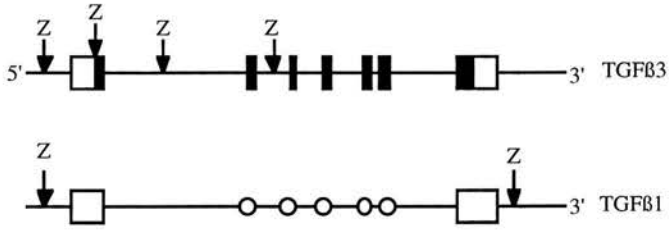


Figure 3.2 Comparison of partial restriction map of the chicken TGFß3 and TGFß1. Solid or open box/bar: exons; circle: undetermined exon; arrow: restriction site; Z: XhoI.

3.3 Screening Chicken Genomic Library

Three TGFß1-specific synthetic oligonucleotide probes (I: 5'TGGTCTCCATGGGG ATCCGCCGCAGCTCTTTGG 3'; II: 5' GCGCCGAGCTGAGGATGCTGCGGC AAAAGGCGG 3' and III: 5' TCCAATATGGTGGTCCGTGCCTGCAAGTGCAGC 3'), and a cDNA probe (nucleotides 49-277) were used to screen a White Leghorn chicken genomic library (library I). A total of 2 x 10^6 phage plaques were screened with each probe. Six, five and four positive plaques were identified for oligo probes I, II and III respectively, together with two positives for the cDNA probe. DNA from these positive clones was subsequently subjected to restriction digestion with either *HindIII* or *SmaI*. Based on the restriction patterns, we were able to classify these exon-specific positive clones into several groups, as shown Figure 3.3.

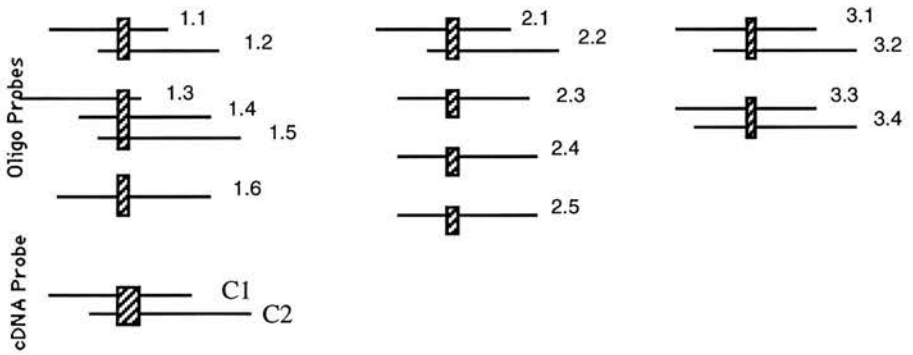


Figure 3.3 Clones obtained from screening of chicken genomic library. Shaded rectangle: exon region; lines connected by shaded bar: overlapping clones.

Subsequently, DNA from each positive clone was digested with a series of restriction enzymes: *BamHI, BglI, EcoRI, HindIII, KpnI, SalI, SmaI, SstI and XhoI*; electrophoresed on 1% agarose gel with λ DNA markers. The sizes of the restriction fragments were calculated from the relative migration ability with the known size markers. Then crude restriction maps were determined. Following electrophoresis, gels were blotted to nylon membrane filter, and probed with [γ^{32}p]ATP or

[α^{32}p]dCTP labeled oligo or cDNA probes. Weak hybridisation was found for oligo I specific clones (i.e. 1.1, 1.2, 1.3, 1.4, 1.5 and 1.6) and oligo II specific clones (i.e. 2.1, 2.2, 2.3, 2.4, and 2.5). Strong hybridisation was revealed for oligo III specific clones (i.e. 3.1, 3.2, 3.3 and 3.4). The filters were stripped and probed with [α^{32}p]dCTP labeled full length TGFß1 cDNA probe. As shown in Figure 3.4, clone 3.3 and 3.4 showed strong hybridisation to TGFß1 whole cDNA probe, whereas clone 3.1 and 3.2, which showed hybridisation to oligo probe III, did not hybridise with this TGFß1 cDNA.

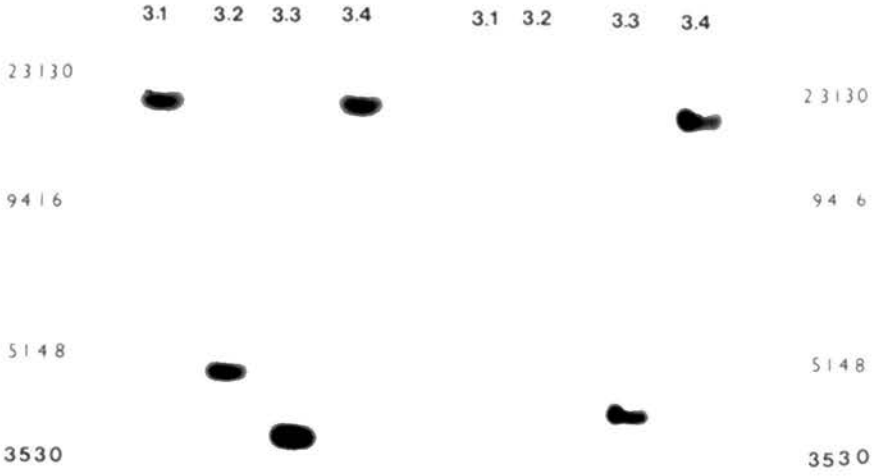


Figure 3.4 Southern blot analysis of genomic clones. (A) Probed with oligo probe III; (B) Probed with TGFß1 whole cDNA probe.

Genomic clones 3.1 and 3.3 were chosen for further studies. The positive bands identified by Southern hybridisation were subcloned. Recombinant plasmid DNAs were digested with the restriction enzymes, *Sau3A* and *PstI*. Refined restriction maps were determined. Finally the positive fragments were subcloned into M13mp18 vectors. DNA sequences were generated using the Sequenase and Taquenase systems (USB). The nucleotide sequence information was subjected to a FASTA search of the

EMBL database. Figure 3.5 showed that across a 300 bp sequence of clone 3.3, there exists a 99% identity with the 3' coding region of chicken TGFß3 cDNA. Therefore genomic clones 3.3 and 3.4 were derived from chicken TGFß3 genomic fragments. However, no homology was obtained with the sequences in the database using the sequence of clone 3.1.

To summarise, a chicken genomic library was screened with a chicken TGFß1 cDNA probe, and chicken TGFß1 specific oligonucleotide probes. A number of overlapping clones were obtained. However, further characterisation of these clones demonstrated that they were not chicken TGFß1 genomic fragments. Two overlapping clones (3.3, 3.4) obtained from the library screening with oligonucleotide probe III have been shown to be chicken TGFß3 genomic fragments.

```
1731 CCTGCAGGAGGGAAAACCTTCGTAATTCTTCAGGCCAGTCGGGCGGGCAG 1682
     ||||||||||||||||||||||||||||||||||||||||||||||||||
  14 CCTGCAGGAGGGAAAACCTTCGTAATTCTTCAGGCCAGTCGGGCGGGCAG 63

1681 AGGTCCCAGGGAAGCTACAGCAAGAAGGGGAAACACTGCCCCATTTCCAA 1632
     |||||||||||||||||||||||||||||||| ||||||||||||||
  64 AGGTCCCAGGGAAGCTACAGCAAGAAGGGGAAACAATGCCCCATTTCCAA 113

1631 TCCCTCTTTTTAACTGCTCTTTCTCATTCCTTGCCTTCCCAGTTCAACTA 1582
     ||||||||||||||||||||||||||||||||||||||||||||||||||
 114 TCCCTCTTTTTAACTGCTCTTTCTCATTCCTTGCCTTCCCAGTTCAACTA 163

1581 CCCCAGCATTCAGATACACAGCAGTTCCTTCACTGTCCTTCTCTTCCGCA 1532
     |||||||||||||||||||||||||||||||| ||||||||||||||||
 164 CCCCAGCATTCAGATACACAGCAGTTCCTTCACTCTCCTTCTCTTCCGCA 213

1531 TCAACTGTCCAGTCATTTTGGAGTGTGGTCTTTGTGTCGTCAAAGCAACT 1482
     ||||||||||||||||||||||||||||||||||||||||||||||||||
 214 TCAACTGTCCAGTCATTTTGGAGTGTGGTCTTTGTGTCGTCAAAGCAACT 263

1481 CCCAGAGGGAACGCGACACTGGCAACAAGACCGAGAA 1445
     |||||| ||||||||||||||||||||||||||||||
 264 CCCAGA.GGAACGCGACACTGGCAACAAGACCGAGAA 299
```

Figure 3.5 Sequence alignment of clone 3.3 (lower) and chicken TGFß3 cDNA upper). Sequence comparison was carried out using the BESTFIT program of the GCG package.

3.4 Conclusions and Discussion

The chicken TGFß1 gene exists as a single copy with an upper size limit of approximately 15 to 22 kb. However, screening of a lambda phage chicken genomic library using both cDNA and oligonucleotide probes failed to obtain any chicken TGFß1 clones. A few possibilities might contribute to this phenomenon.

1) Low probe specificity

First, all the synthetic oligo probes used were 30 to 33 bp long. The specificity of these probes may be low. Second, Burt *et al* have shown that TGFß1 and TGFß3 are more related to each other than to TGFß2 (Burt and Paton, 1992). A comparison of their cDNA sequences revealed that the 3' region of the two genes were highly homologous. The oligonucleotide probe III is localised within this homologous region. Therefore oligonucleotide probe III may pick up TGFß3 genomic clones from the chicken genomic library. Third, the TGFß1 cDNA was very GC-rich. The average GC content in the TGFß1 cDNA was 66%, while the 5' coding region has a GC content over 70%. Additionally, several sites within the cDNA include stretches of poly-[C] or poly-[G]. This unusual GC content may cause the formation of stable secondary structures and would be one aspect contributing to the low specificity of the probe.

2) Absence of TGFß1 genomic fragments in the chicken genomic library

Palazzolo (1991) pointed out that some regions of the genome are not represented in genomic libraries because of the relative difficulty of cloning certain genomic regions. Generally in a three-hit library, almost 5% of the genomic sequences would be uncloned; in a four-hit library about 2% and in a five-hit about 1%. Additionally, high GC-content sequences together with its methylated state would make the passage

of replication forks difficult during phage or plasmid replication (Hansen, 1993). Furthermore, the first chicken genomic library used was an amplified Lambda EMBL3 phage library. For these reasons, DNA fragments with high GC content such as TGFß1 might not be replicated in phage, or might be replicated very poorly and therefore may have been lost during library amplification. An alternative explanation may be due to the formation of stable secondary structures, with the consequence that sequences may not be amplified at all. Support for this view comes from attempts to PCR regions of the TGFß1 gene. It was found that two reverse primers (21-mer) which are 60 bp apart from each other in the TGFß1 5' coding region showed significantly different PCR efficiencies with a common forward primer. The longer PCR product which contained a stretch of poly-$[C/G]_{32}$ could only be amplified using deaza dCTP which helps to overcome the problems associated with sequences with high GC content and ability to form stable secondary structures.

To check the representativeness of the library, another chicken genomic library (Strategene) was screened using the chicken TGFß1 cDNA probe. As a positive control, chicken ornithine decarboxylase (ODC) cDNA probe was used to screen for chicken ODC genomic clones. A total of eight positive ODC clones, whereas no TGFß1 positive clone were obtained. This result indicated that TGFß1 genomic fragments were not represented in these chicken genomic libraries regardless of their existence in the chicken genome.

Chapter 4

Construction of Microsatellite-Enriched Genomic Library by Genetic Marker Selection

4.1 Introduction

"Hybridisation screening" protocols have been used successfully in isolating microsatellites from human and other mammalian genomes (Dietrich *et al.*, 1992; Weissenbach *et al.*, 1992; Rohrer *et al.*, 1994). The chicken genome is estimated to contain 10^4 $[CA/TG]_n$ microsatellites. This is approximately 10-fold less than that of the human genome (Spada *et al.*, 1991; Koide *et al.*, 1994). Consequently, a random small-insert chicken genomic library with an average insert size of 300-500 bp would be expected to have a frequency of $[CA/TG]_n$ repeats of less than 0.2%. This makes the hybridisation screening method inefficient in isolating microsatellites from the chicken genome. Furthermore, this approach does not allow intensive concentration on a single region and becomes progressively less efficient as saturation is approached. Therefore it is desirable to develop a method for enriching microsatellites in chicken genomic libraries. In an initial attempt to develop such a method, a "genetic marker selection" method was adapted (Ostrander *et al.*, 1992). Using this method, a 12-fold and 8-fold enrichment of microsatellites were obtained in $[CA/TG]_n$ and $[CCT/AGG]_n$ selected genomic libraries.

4.2 General Frequency of Microsatellites in the EMBL Database

The EMBL database was searched for various types of di-, tri- and tetra-nucleotide tandem repeat-containing sequences using the PATTERN/SEARCH program of the GCG package. The repeats for pattern-searching were restricted to 20 bp or longer. The results of this search are shown in Figure 4.1.

Previous studies (Beckman and Weber, 1992; Moran, 1993) along with the results of this database search showed that $[CA]_n$ and $[CCT]_n$ are most frequent. Therefore two

synthetic oligonucleotides, $[CA]_{15}$ and $[CCT]_{10}$ were synthesised and used in genetic marker selection.

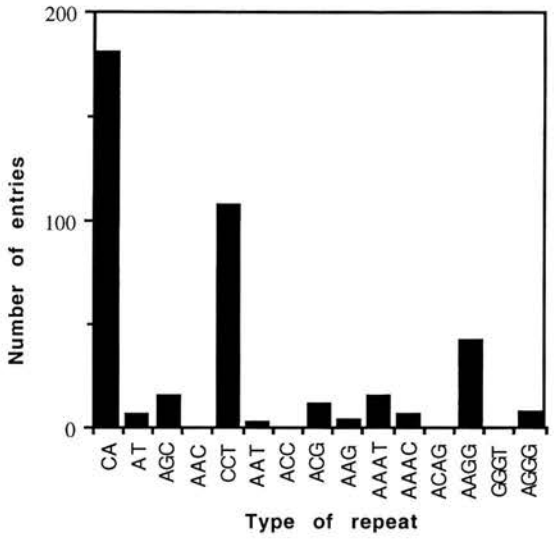


Figure 4.1 General frequency of simple sequence repeats in the EMBL database. The type of repeat was plotted against the number of entries.

4.3 Enrichment of Microsatellites by Genetic Marker Selection

The procedure is shown in Figure 4.2. Chicken genomic DNA was digested to completion with restriction enzymes *RsaI* and *HaeIII* . Restricted DNA was size selected (approx. 150-500 bp) on 1% low melting point agarose gel and subsequently cloned into the *SmaI* site of the phagemid vector, pTZ19U.

The ligation mixture was subjected to restriction digestion with *SmaI*. Recombinant molecules have lost the *SmaI* cloning site and remain circular. The circular recombinant molecules transform *E. coli* cells at much higher frequency than the linearized vectors. This reduced the background of non-recombinant colonies. The ligation mixture was transformed into *E.coli* strain, CJ236 which carries *dut* and *ung* mutations. When DNA is synthesised in a *dut ung* double mutant bacterium, the

nascent DNA carries a number of uracils substituted for thymine as a result of the *dut* mutation. This mutation inactivates the enzyme dUTPase and results

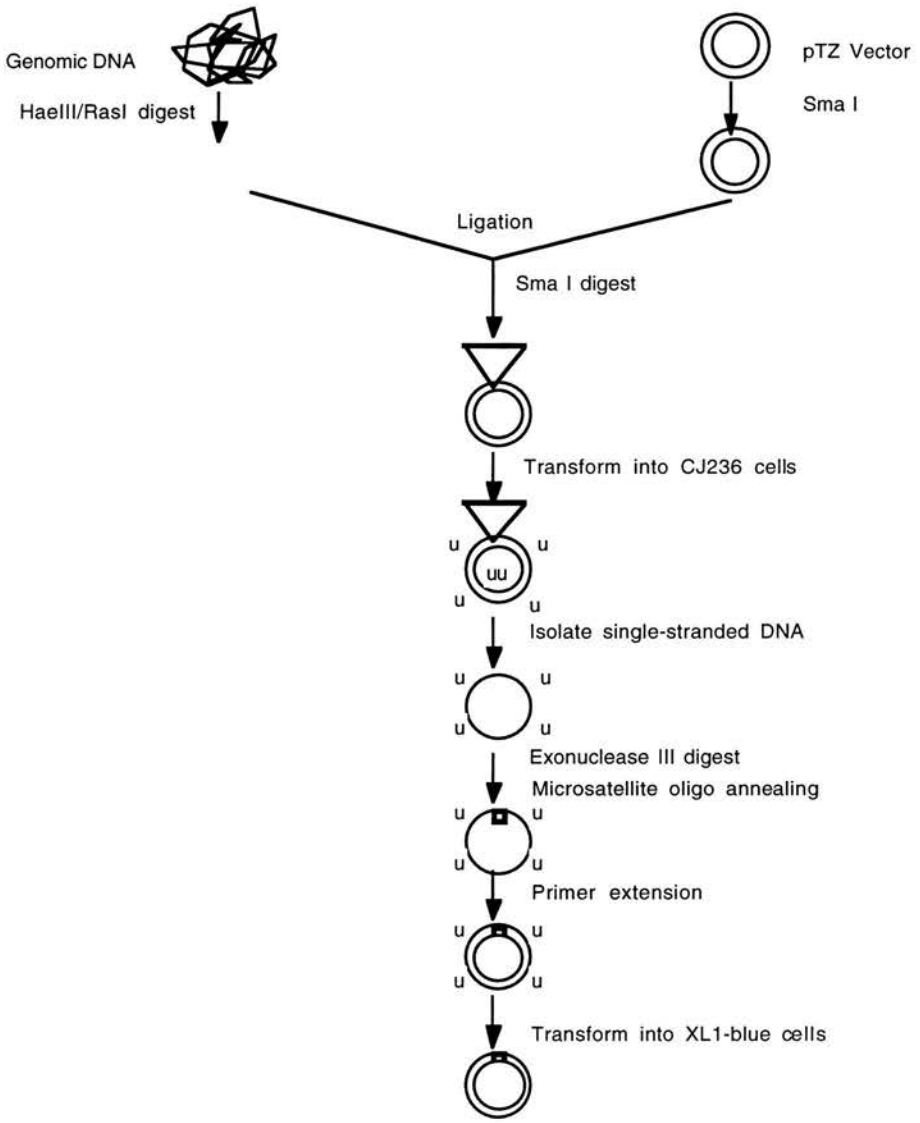


Figure 4.2 Construction of microsatellite-enriched genomic DNA library by genetic marker selection. U: uracil; solid box: microsatellite primer.

in high intracellular levels of dUTP. The *ung* mutation inactivates uracil N-glycosylase, which allows the incorporated uracil to remain in the DNA (Ostrander *et al.*, 1992).

Following the method described in section 2.2.15 of Materials and Methods, uracil-rich single-stranded phagemid was harvested. The phagemid stock obtained was estimated to contain 4 x 10^{11} pfu/ml by transfection of CJ236 cells. The single-stranded DNA was extracted and subjected to exonuclease III (Promega) digestion to remove any linear DNA molecules remaining in the single stranded circular DNA templates. This was necessary to prevent such linear molecules acting as random primers during primer extension reaction.

Microsatellite oligonucleotides $[CA]_{15}$ or $[CCT]_{10}$ were phosporylated at the 5' end and annealed to the single-stranded DNA template. The single stranded template was then converted into a double stranded circular DNA molecule *in vitro* using Taq DNA polymerase and the microsatellite oligonucleotide as a primer (M&M 2.2.15). The extended DNA was transformed into XL1-Blue cells by electroporation (M&M 2.2.8). Since XL1-blue contains wild type *dut* and *ung* genes, any DNA strands which have incorporated uracil will be destroyed by the *ung* enzyme. In principal, only those DNA strands resulting from primer extension will survive to replicate in XL1-blue cells. These DNAs should contain microsatellite sequences. This procedure should therefore provide an enriched pool of plasmids containing the target microsatellite sequence.

4.4 Enrichment of Microsatellite Clones in the Selected Libraries

The marker-selected small insert genomic libraries were screened using the method described (M&M 2.2.17) with [γ^{32}p] ATP labeled microsatellite oligonucleotide

probes. Positive clones were identified by autoradiography. As shown in Figure 4.3, when comparable numbers of colonies were screened, positives were only detected in the marker-selected genomic library. No signal was detected in the unselected control library. This result indicated that, through the procedure of marker selection, the proportion of microsatellite-containing clones had been increased.

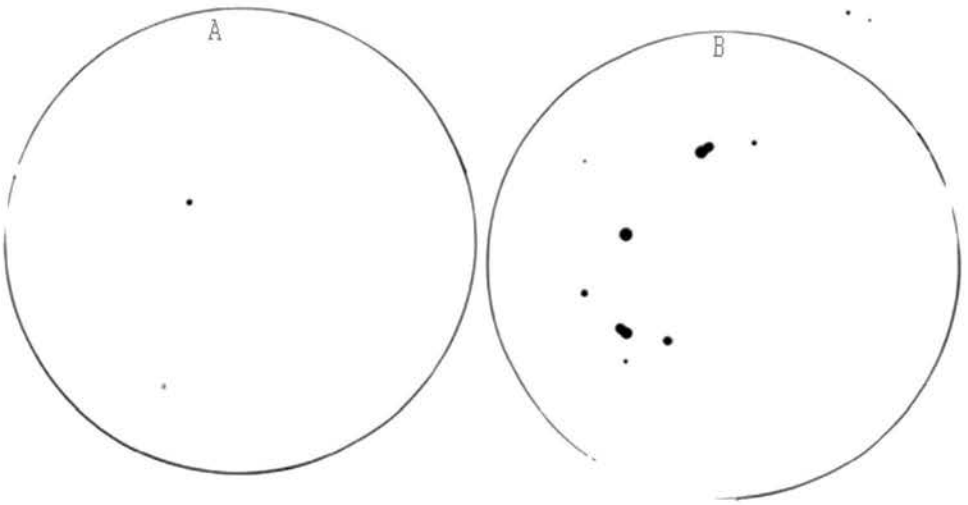


Figure 4.3 Enrichment of $[CA/TG]_n$-containing clones in genomic DNA libraries. Libraries were constructed using the marker-selection protocol, Three hundred clones were screened with [γ^{32}P] ATP labeled $[CA]_{15}$. (A) unenriched; (B) $[CA/TG]_n$-enriched.

The percentage of positive clones in the marker-selected genomic library was calculated and is listed in Table 4.1

Table 4.1 Comparison of numbers of positive clones in the marker-selected and non-selected genomic DNA libraries

	Positive Clone Content	
Marker	**$[CA]_n$**	**$[CCT]_n$**
Marker-Selected	1.5 (%)	0.5 (%)
Unselected	0.12 (%)	0.06 (%)
Enrichment	12.5 (fold)	8.3 (fold)

Theoretically a higher proportion of the colonies in the marker selected libraries should contain the selected microsatellite. However, in practice, only a small proportion of the colonies screened (0.5 to 1.5%) were positive depending on the specific microsatellite oligo used. The percentage of positive clones in our $[CA]_{15}$-selected library, 1.5%, is quite close to that reported by Cheng *et al.* (Cheng and Crittenden, 1993) with chicken genomic DNA using marker selection. However, this figure is much lower than that obtained with mammalian genomic DNA, where as many as 50% of the clones were reported to be positive in the marker-selected library (Ostrander *et al.*, 1992). The "background" negative clones might be generated in three possible ways.

1) Contamination by double-stranded phagemid DNA;
2) Random priming by short, linear DNA molecules remaining in the single-stranded template DNA (e.g. *E.coli* chromosome DNA and RNA etc.) rather than the specific primer;
3) Products of random priming by secondary structures within the single stranded DNA molecule itself.

The first possibility seems unlikely since the treatment of the single stranded DNA template with a restriction enzyme, *HindIII*, did not increase the percentage of positive clones. Exonuclease III cleavage of the DNA template, although resulting in a ten-fold reduction of the number of total colonies, also did not solve the problem.

This indicated that either incomplete cleavage of random primers by Exonuclease III or priming of the extension reaction by secondary structures within the single stranded template itself was responsible for the appearance of the high number of negative clones.

4.5 Sequence Characterisation of Positive Clones

DNA was prepared from positive clones using the Qiawell-18 plasmid purification kit (M&M 2.2.9). *EcoRI* and *HindIII* double digestion of plasmids was carried out to check the quality and the size of the inserts. Ninety five percent of the positive clones have an insert of 150-500 bp in length. The remaining 5% of the positive clones have either a large insert (>500 bp) or could not be digested by the restriction enzymes. Nucleotide sequence information was generated using the Sequenase or the Taquenase systems (USB). Sequences were edited and analysed using the *GCG* package (University of Wisconsin Genetic Computing Group) or *Seqman* (DNA Star). Figure 4.4 shows examples of DNA sequences from positive clones.

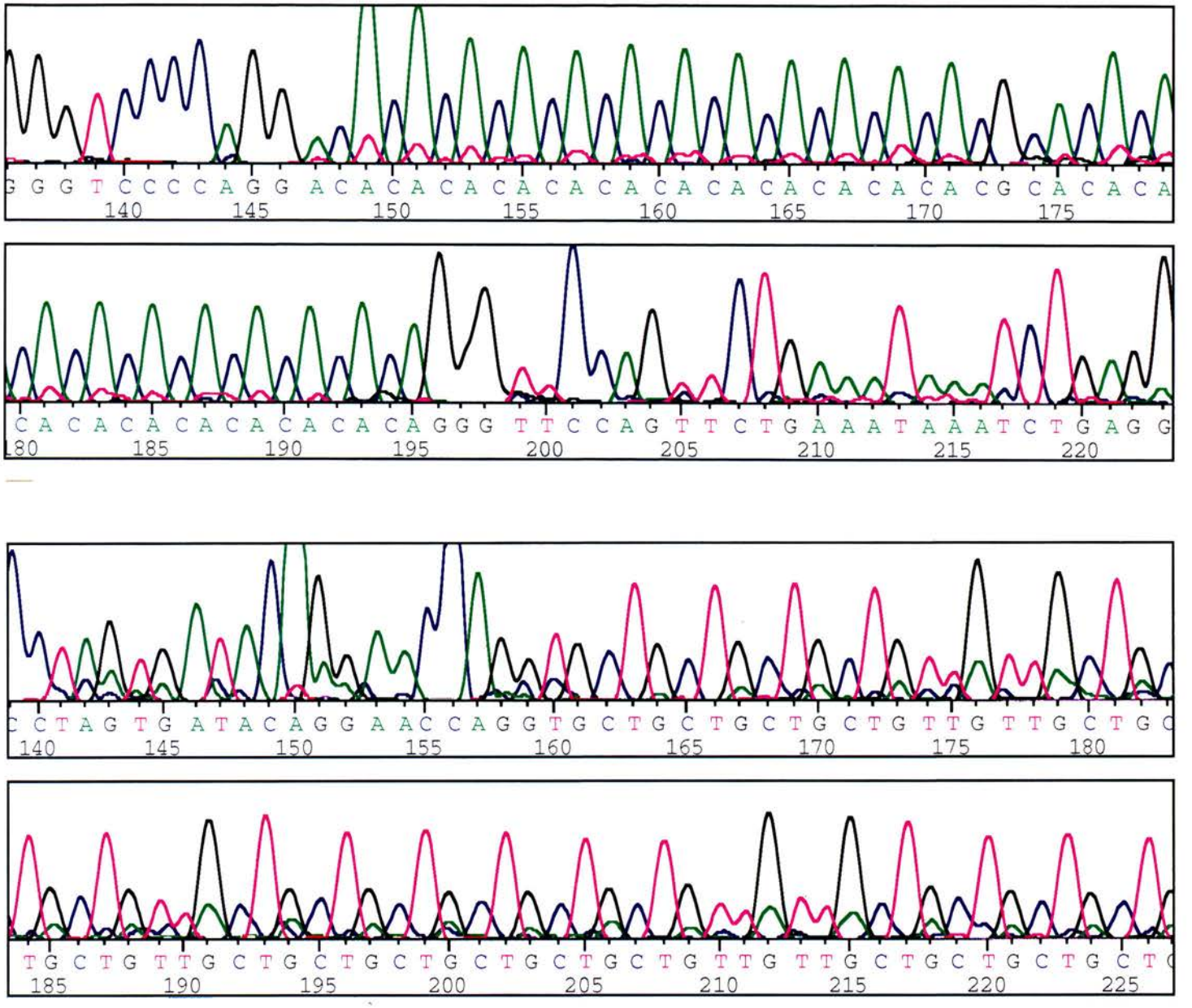


Figure 4.4 Examples of DNA sequences of microsatellite clones from marker-selected genomic DNA libraries. A. $[CA]_n$ microsatellite; B. $[CTG]_n$ microsatellite.

A total of 21 and 37 positive clones were identified and isolated from the $[AC]_{15}$- and $[CCT]_{10}$-enriched genomic DNA libraries respectively. Further characterisation of these positive clones by sequencing revealed 7 unique contigs for $[AC]_n$-type clones, and 14 unique contigs for $[CCT]_n$-type clones. The nucleotide sequence compositions of the repeat motif for each contig are shown in Table 4.2 and Table 4.3.

Table 4.2 DNA sequence of the repeat motif of positive clones from the $[CA/TG]_n$-enriched genomic library

Contig	Repeat Motif	Number of Clones
1	$[AT]_8[CA]_7$	1
2	$[AT]_6[GT]_{13}$	1
3	$[AC]_{17}$	1
4	$[AC]_{18}$	3
5	$[AC]_{19}$	1
6	$[AT]_{10}[CA]_{19}$	6
7	$[AC]_{28}$	1
	Anonymous	7

4.6 Conclusions

(1) About half of the positive clones picked from the $[CA/TG]_n$-enriched library do not contain any of the target repeats. This may have been due to the low stringency during colony hybridisation. To overcome this, either a higher stringency hybridisation or rescreening step should be carried out.

(2) Of the microsatellite-containing clones from the $[CA/TG]_n$-enriched library, the [CA] or [TG] repeats were usually long (repeating 17 to 28 times). About half of these clones contain single perfect [CA]-repeat. The rest of the clones show complex repeats, for example, often two distinct types of perfect repeats adjacent to each other.

Table 4.3 DNA sequence of repeat motifs of positive clones from the $[CCT]_n$-enriched genomic library.

Contig	Repeat Motif	Number of Clones
1	$[CCT]_2$	1
2	$[CCT]_3$	1
3	$[CCT]_3$	1
4	$[CCT]_3[CT]_{17}$	1
5	$[CCT]_4$	1
6	$[CCT]_4$	1
7	$[CCT]_4$	1
8	$[CCT]_4$	2
9	$[CCT]_5$,AG-rich	1
10	$[CCT]_5$,CT-rich	1
11	$[CCT]_5[AAT]_7$	2
12	$[CCT]_5[AAT]_7$	1
13	$[CCT]_5[TTTA]_4$	1
14	$[CCT]_6$	1
	[CCT]-rich	12
	Anonymous	9

(3) Of the 37 positive clones from the $[CCT]_n$-enriched library, 9 clones do not contain any [CCT] repeat (24%), 12 clones are CT-rich sequences (32%). The [CCT] repeats in the rest of these clones are usually short (repeating 2 to 6 times). Such short tandem repeats are expected to be less polymorphic.

(4) As shown in Table 4.2 and 4.3, in both the $[CA]_n$ or $[CCT]_n$-enriched genomic libraries, a number of over-represented clones were found. The small-insert libraries are therefore not representative of the genome.

This method seems impractical for generating microsatellite-enriched genomic libraries necessary for genome mapping experiment in chicken. A new method is required.

Chapter 5

Construction of Microsatellite-enriched Chicken Genomic Libraries By DNA Affinity Hybridisation

5.1 Introduction

Although genetic marker selection has been shown to be an efficient method for enriching specific microsatellite in a small insert genomic library, we encountered a few problems with this method.

1) The percentage of positive clones in the libraries was still low. Screening of the $[CA/TG]_n$ marker-selected genomic library showed that only about 1.5% of the clones were positive clones.
2) During the *in vitro* primer extension step of the procedure, primer mismatches may occur. Therefore there is a high risk of getting mutated sequences in using this method.
3) The biggest problem was the low complexity of the marker-selected library. Both in our hands and others (Cheng and Crittenden, 1993), it was found that a number of clones were over-represented. This could be a reflection of the incomplete genome coverage used in the marker-selected method. Assume that the average size of the inserts was 350 bp (150-500 bp) and the chicken genome to be 1.2×10^9 bp in size. Since the marker-selection procedure started with a total of 5×10^4 recombinant clones, the maximal genome coverage of these clones was about 1% ($3.5 \times 10^7 / 1.2 \times 10^9$). Because the chicken genome contains roughly 10^4 $[CA/TG]_n$-type microsatellites, about 100 distinct types of positive clones are expected in the starting-library. It was assumed that all microsatellites would be perfect [CA] repeats and therefore all could be selected by the primer. However, more than half of the microsatellites in the chicken genome belong to imperfect [CA] repeats. These microsatellites would not be selected by the marker-selection procedure unless significant "mismatches" are allowed. This would reduce the total number of microsatellites generated in the marker-selected library to about 50.

To avoid these limitations of genetic marker selection, an alternative method is needed for (1) enriching for microsatellites through physical rather than biological process so that the enriching conditions will be easier to control and the probability of getting mutated sequences be reduced; (2) leaving DNA cloning until the very last step, i.e. after microsatellite-enrichment, so that the total number of colonies to be manipulated will be kept to a minimum.

A PCR-based DNA affinity hybridisation and capture procedure was developed for the construction of microsatellite-enriched genomic DNA libraries. In this procedure, a 5'-biotinylated $[CA]_{15}$ oligonucleotide was used to hybridise with heat-denatured $[CA]_n$ microsatellite-containing DNA fragments. The heteroduplex molecules were then physically captured and separated from non-hybridised DNA molecules by streptavidin-coated magnetic beads. The captured DNA was PCR amplified and cloned into a plasmid vector to produce a genomic library.

This method has a number of advantages over the marker-selection procedure. First, the DNA affinity technique allows a wide range of mismatching base pairs between the oligonucleotide and DNA fragment. Second, the degree of such mismatches can be controlled by the conditions of hybridisation and the stringency of washing. so that this method should select more longer repeats. These kinds of microsatellites may be more polymorphic. Thirdly, the method is flexible, for example, it can start with a very small amount of genomic DNA, and the affinity-hybridisation step can be repeated several times. Finally, one of the lone-linker oligonucleotides was used as primers in PCR amplification, this should increase the accuracy of primer annealing and avoid modified clones induced from the primer mispairing in the marker selection procedure.

5.2 DNA Affinity Hybridisation/Capture Protocol for Enriching Microsatellites

The protocol employed is shown in Figure 5.1, chicken genomic DNA was randomly fractionated by sonication. The sonicated DNA was size-selected (150-500 bp) on a 1% gel using low melting-point agarose. The DNA was made blunt-ended by end-filling using T_4 DNA polymerase. In order to facilitate PCR amplification of the DNA, 'lone-linkers' generated from the annealing of two partially overlapping oligonucleotides (M&M 2.2.16) were ligated to the ends of the size-selected genomic DNA by T_4 DNA ligase. The lone-linkers were specially designed in order not to allow ligation between the linkers. An *EcoRI* restriction site was included on the linker to facilitate subsequent cloning into a plasmid vector. An aliquot of the linker-attached DNA (1-2 ng) was amplified by polymerase chain reaction (M&M 2.2.5).

The amplified DNA was denatured by heating in a boiling-water bath for five minutes followed by quick chilling in an ice-bath. The denatured genomic DNA was mixed with an aliquot of the 5'-biotinylated $[CA]_{15}$ oligonucleotide. The mixture was heated to 95^0C for 1-2 minutes, then transferred to a pre-set 60^0-70^0C water bath for 10 minutes followed by chilling in an ice-bath. Pre-washed streptavidin-coated magnetic beads were added. The affinity hybridisation was carried out at room temperature for one hour in a total volume of 200 µl with 0.1% SDS, 0.1 M NaCI, 10 mM Tris-HCI (pH 8.0) and 0.1 mM EDTA. During this procedure, DNA fragments that contain the target repeat sequence will hybridise with the 5'-biotinylated microsatellite oligo. Streptavidin-coated magnetic beads, to which the 5'-biotinylated oligo binds, were captured by placing the tube on a magnetic stand. The matrix was successively washed with buffer A five times at room temperature, and once at 70^0C; followed by two washes with buffer B at 70^0C (M&M 2.2.16).

As a consequence, all DNA fragments except those bound by the 5'-biotinylated oligo were washed away. The bound DNA was eluted and amplified by PCR. The amplified DNA was proteinase K treated and digested with the restriction enzyme, *EcoRI* and subsequently cloned into the *EcoRI* site of the plasmid vector, pBluescript(+). The ligation reaction was optimised to maximise the number of recombinant clones and the ligation mixture was transformed into XL1-Blue cells by electroporation (M&M 2.2.3). Finally the enriched library was titrated by plating onto

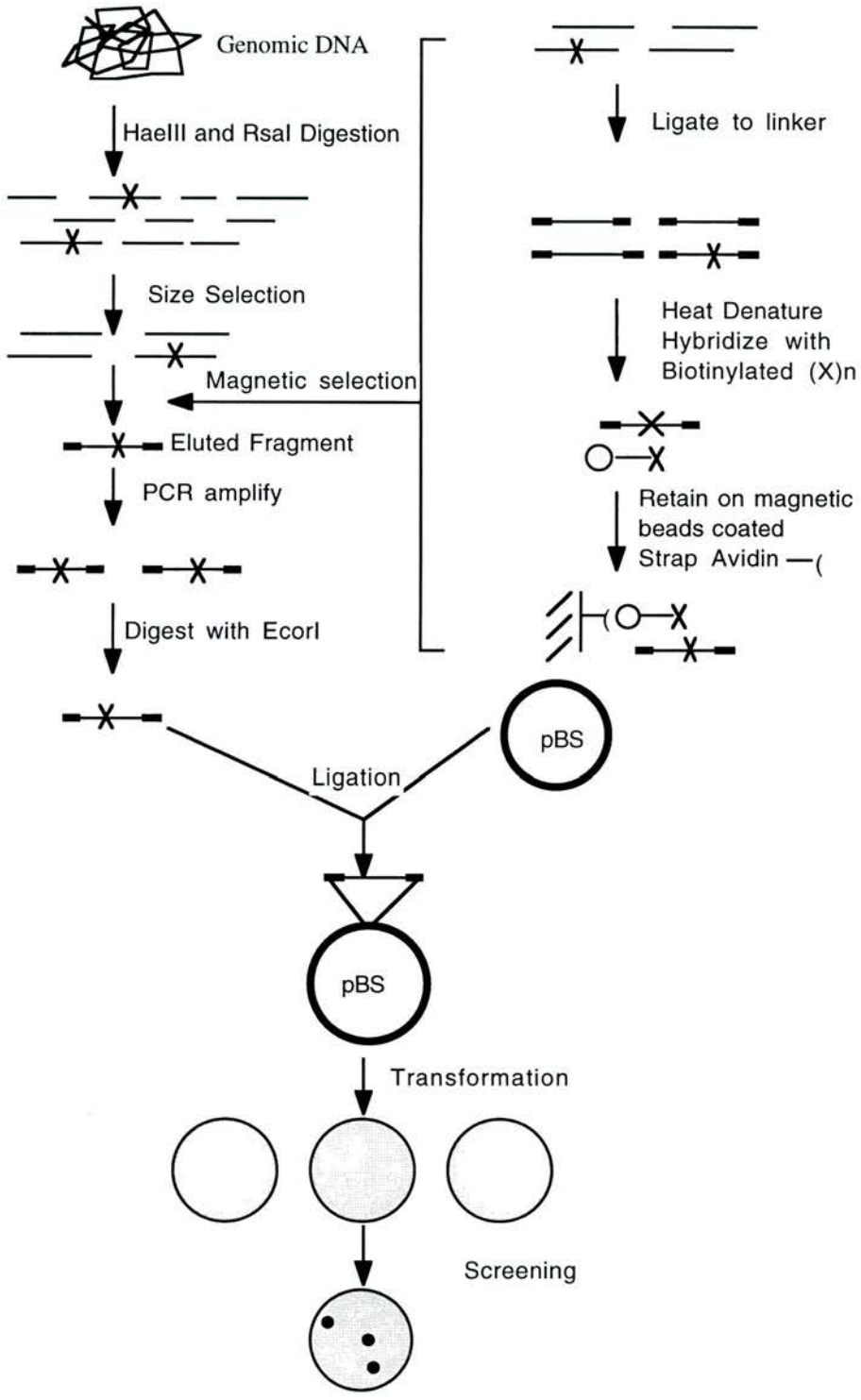


Figure 5.1 Schematic of the DNA affinity hybridisation protocol for constructing microsatellite-enriched genomic libraries. Crosses indicate microsatellites; black dots indicate positive clones.

X-gal-ampicillin plates. The number of recombinants indicated by white colonies was usually about 70%.

5.3 Enrichment of $[CA/TG]_n$ Microsatellites in the Selected Genomic Library

The enriched library described above was screened once with [γ^{32}P] ATP labeled $[CA]_{15}$ oligo nucleotide probe. As shown in Figure 5.2, when comparable numbers of colonies were screened using this probe, positive clones were only detected in the $[CA/TG]_n$-enriched genomic library, but not in the unenriched control library. This result indicated that the affinity hybridisation-selected chicken genomic library was highly enriched in $[CA/TG]_n$ microsatellites.

A total of 5,000 colonies from both the control and the $[CA/TG]_n$-enriched library were screened, the percentages of positive clones were calculated and listed in Table 5.1. The percentage of positive clones was 5.77% in the $[CA/TG]_n$-enriched genomic library, whereas in the unenriched control library it was only 0.12 percent. This result indicated that using the DNA affinity hybridisation selection procedure, an approximate 50-fold enrichment of $[CA/TG]_n$-type microsatellites had been achieved.

Table 5.1 Comparison of percentages of positive clones in the $[CA/TG]_n$-enriched and unenriched genomic DNA libraries

DNA library	**% of Positive Clones**
Unenriched	0.12
Enriched	5.77
Enrichment	50 (fold)

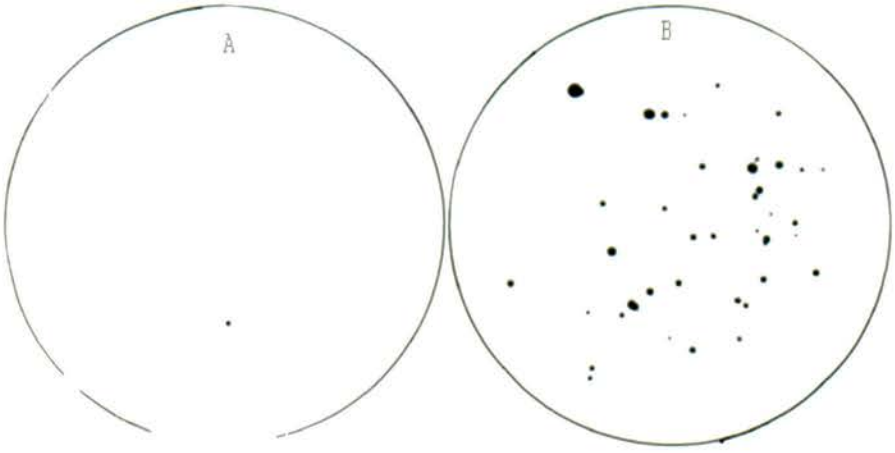


Figure 5.2 Enrichment of $[CA/TG]_n$ microsatellites in genomic DNA library. Libraries were constructed using DNA affinity hybridisation and capture protocol. Three hundred clones from (A) Unenriched control; (B) $[CA/TG]_n$-enriched were screened with $[\gamma^{32}P]$ATP labeled $[CA]_{15}$ oligo probe.

5.4 Sequence Characterisation of the Positive Clones from the (AC)-enriched Library

After screening, positive clones were lifted and propagated in LB broth. Plasmid DNA was prepared using the Qiawell-18 Plasmid DNA Purification Kit (M&M 2.2.9). The quality, quantity and the size of the inserts were checked by digestion with the restriction enzyme, *EcoRI*, and electrophoresis on a 1% agarose gel. Then clones with inserts were selected for further characterisation by DNA sequencing. Nucleotide sequences of microsatellite clones were generated using the Taq Cycle Sequencing System (ABI) using either T3 or T7 primers. The products were analysed using an ABI 373 automated DNA sequencer. Figure 5.3 shows two examples of DNA sequences from positive clones.

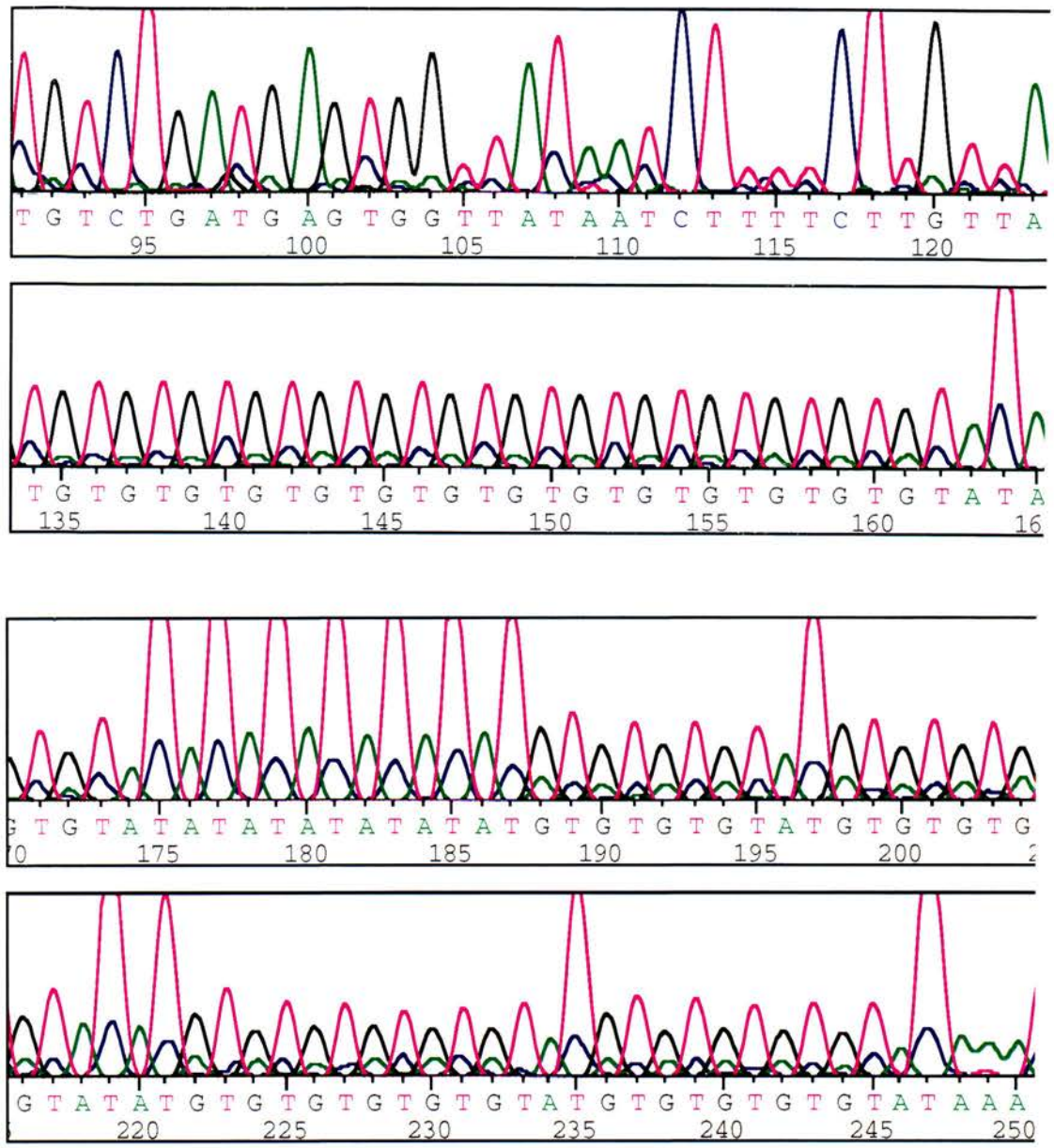


Figure 5.3 Examples of DNA sequence of microsatellite clones from the genomic library constructed using the DNA affinity hybridisation protocol. (1) Perfect [CA/TG] repeat; (2) Imperfect [CA/TG] repeat.

A total of 51 positive clones have been characterised by DNA sequencing. The sequence motifs found in these clones are given in Table 5.2.

(1) General characteristics of the $[CA]_n$-enriched library

A. The percentage of positive clones in the enriched library was calculated to be 5.77%, whereas in the non-selected library it was 0.12%. This indicates an

approximate 50-fold enrichment of $[CA/TG]_n$ microsatellites in the library. It is also about a 4-fold enrichment over the marker-selected genomic library described in Chapter 4.

B. Further characterisation of these positive clones by DNA sequencing showed that about 87% of these positive clones were real $[CA/TG]_n$-containing DNA sequences.

Table 5.2 Sequence characterisation of repeat motifs of positive clones from the $[CA/TG]_n$-enriched genomic DNA library

Contig	DNA Sequence of the repeat motif	Number of Clones
1	$[AC]_{10}$	1
2	$[AC]_{12}$	1
3	$[AC]_{12}[AT]_6[AC]_{11}[AT]_7$	1
4	$[AC]_{13}$	1
5	$[AC]_{17}$	1
6	$[AC]_{18}$	1
7	$[AC]_{20}$	1
8	$[AC]_{20}$	1
9	$[AC]_{21}$	1
10	$[AC]_{24}$	1
11	$[AC]_{24}$	1
12	$[AC]_{26}$	1
13	$[AC]_{28}$	1
14	$[AC]_{32}$	1
15	$[AC]_{32}$	1
16	$[TCACAC]_{10}$	1
17	$[CACACACAG]_{29}$	1
18	$[TGACACACACA]_4$	1
	[AC]-rich	26
	No $[AC]_n$-containing sequences	7

C. However, approximate half of the real positive clones identified belong to the group of [CA]- or [TG]-rich sequences. Characteristically, in such sequences, instead of one or two tracts of perfect $[CA]_n$ or $[TG]_n$ repeats, long runs of $[CA]_n$ or $[TG]_n$ repeats have been found which were frequently interrupted by unique sequences. The problems with such sequences were: (1) it was difficult to design specific primers for PCR; (2) it indicated possible clustered locations for such sequences in the genome, for example telomere and centromere where simple DNA sequences were usually repeated hundreds to thousands of times (Roger, 1983; Blackburn, 1991).

D. In many clones, the $[CA/TG]_n$ motifs were found at the extreme ends of the inserts. This is probably due to the effect of random breakage of DNA by sonication. These clones are unlikely to be useful because there is not enough flanking sequence to design primers for PCR.

E. As seen from Table 5.2, every clone sequenced so far was unique in sequence composition. This indicates a high degree of complexity in the enriched library.

F. 78% of the real microsatellite-containing clones contain simple [CA] or [TG] dinucleotide repeats, whereas 22% have complex dinucleotide repeats.

(2) Size distribution of $[CA/TG]_n$ motifs in the enriched library

The observed size distributions of $[CA/TG]_n$ motifs in the DNA affinity hybridisation selected genomic library is shown in Figure 5.4. A 'normal distribution' pattern with a modal length of $[CA]_{16-20}$ being the most frequent was observed for the $[CA/TG]_n$ repetitive motifs in the genomic library. This distribution pattern was quite similar to the one observed by Crooijmans *et al.* (1993) in an unenriched chicken genomic library (lambda *ZapII*), where the number of repeats was found varied from 9 to 33 with $[CA]_{13-16}$ being the most frequent motif length. The consistency of the two distribution patterns suggests the validity of the microsatellite-enriched chicken

genomic library constructed in this experiment. However, this distribution pattern was quite distinct from the one observed by Cheng *et al.* (1993) in a marker-selected genomic library, where the numbers of uninterrupted repeats were observed varied from 4 to 14 with the $[CA]_{6\text{-}10}$ being the most frequent motifs.

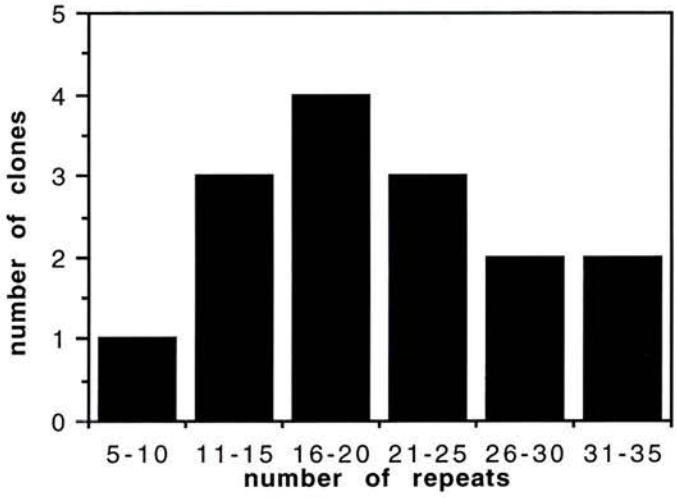


Figure 5.4 Size distribution of $[CA/TG]_n$ repeat motifs of positive clones in the genomic library constructed through DNA affinity hybridisation selection protocol.

Chapter 6

Construction and Utilisation of Microsatellite-enriched Chicken cDNA Libraries

6.1 Introduction

Searches of the current EMBL database revealed that a large number of expressed sequences (cDNA) contain simple sequence repeats ($[CA]_n$, $[CAG]_n$ etc.), among which $[CA/TG]_n$ dinucleotide tandem repeats were found to be the most abundant (Love *et al.*, 1990; Beckman and Weber, 1992; Moran, 1993). Some microsatellites were located within coding regions of genes (mostly triplet repeats), but most of them were localised in non-coding regions, e.g. 5' and 3' untranslated regions.

These microsatellites can be used as markers in linkage studies and have an advantage that they are direct markers of expressed genes. Therefore if a trait was found linked to any of these microsatellite markers, then the associated expressed sequences are likely candidate genes for controlling this trait.

Recently, various functions have been assigned to simple sequence repeats (microsatellites), for example, promoting homologous recombination (Blaho and Wells, 1989) and regulating gene expression (Santoro *et al.*, 1984; Peck, 1985). In addition, since these simple sequence repeats are highly mutable sequence elements, genes containing such sequences may likely be candidate genes to cause disease conditions.

Among various types of simple sequence repeats, $[CAG]_n$ trinucleotide repeats have received much attention recently. $[CAG]_n$ triplet encode poly-glutamine residues and appear more frequently within coding regions of genes rather than the noncoding 5', 3' untranslated regions or introns. Database searches reveal that many eucaryotic transcriptional factors contain $[CAX]_n$ repeats (X=A or G) in their coding regions. The poly-glutamine stretch has been shown to act as a transcription activation/inhibition domain in transcriptional factors. Since these triplets are located

in coding regions, they may be less polymorphic than other microsatellite markers as far as the number of repeats is concerned.

To facilitate the efficient isolation of microsatellites associated with expressed sequences, $[CA/TG]_n$- and $[CAG]_n$-enriched chicken liver cDNA libraries were constructed using the DNA affinity hybridisation and capture procedure developed for genomic sequences.

6.2 Protocol for the Construction of Microsatellite-enriched cDNA Libraries

The procedure employed is shown in Figure 6.1, messenger RNA was extracted directly from chicken liver tissue using the polyA Tract System 1000 (Promega) (M&M 2.2.13). Double-stranded cDNA was synthesised using the Riboclone cDNA Synthesis System (Promega) (M&M 2.2.14). Then the total cDNA was divided and digested separately with restriction enzymes *HaeIII* or *RsaI*. The restriction enzyme digested cDNAs were combined and size-selected (500 to 800 bp) on 1% low melting point agarose gel (M&M 2.2.12). "Lone-Linkers" were ligated to the ends of the cDNA with T_4 DNA ligase. Then an aliquot (1 to 2 ng) of the size-selected linker-attached cDNA was amplified by the polymerase chain reaction (M&M 2.2.15).

The amplified cDNA was then subjected to the DNA affinity hybridisation procedure developed and described in Chapter 5, in which a 5'-biotinylated microsatellite oligonucleotide, $[CA]_{15}$ or $[CAG]_{10}$ was used to hybridise with the cDNA fragments that contains the corresponding tandem nucleotide repeats.

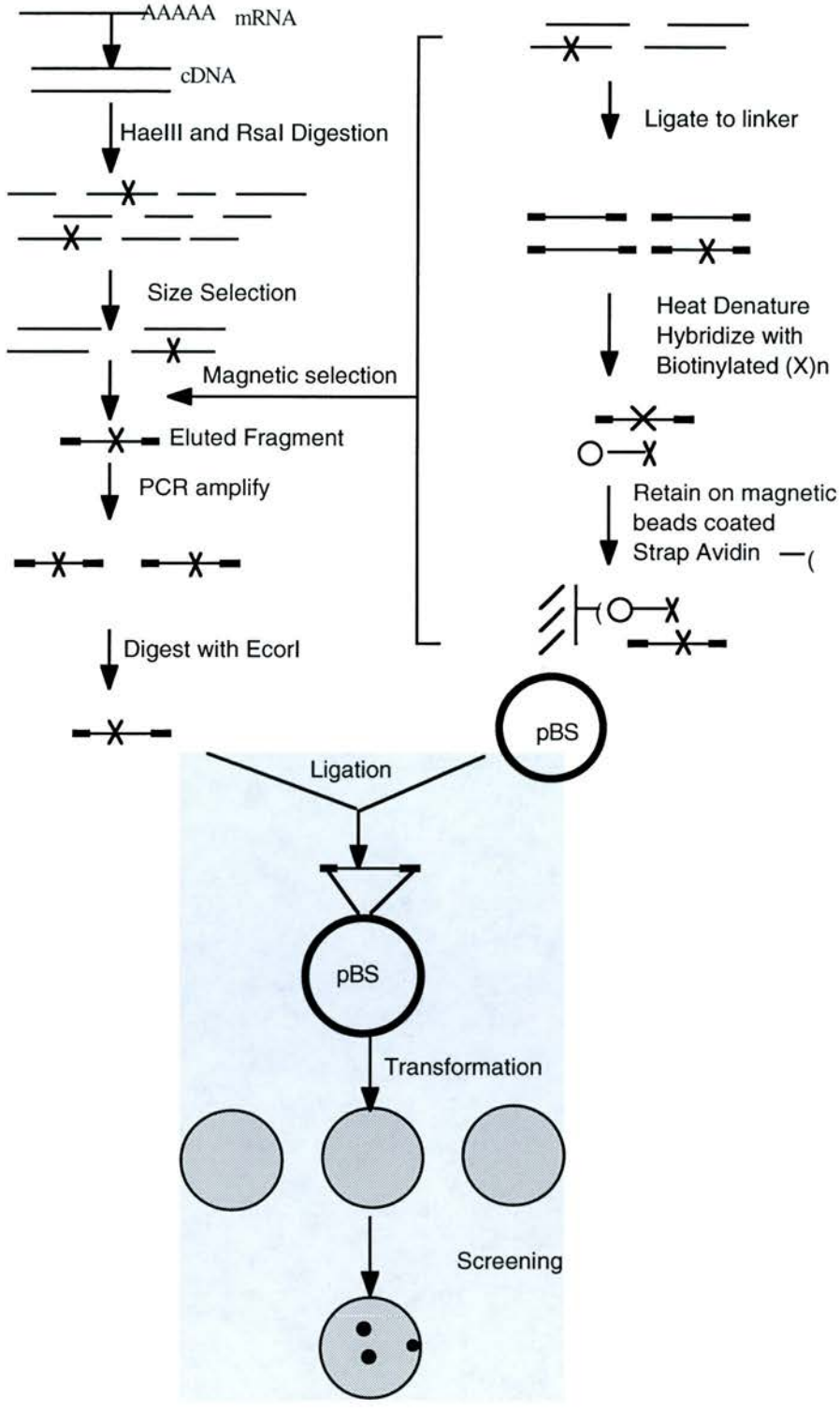


Figure 6.1 Schematic of the construction of microsatellite-enriched cDNA library through DNA affinity hybridisation and capture procedure. Black dot: positive clone; X: simple sequence repeat.

6.3 Characterisation of the [CA/TG]$_n$-Enriched cDNA Library

(1) Enrichment of [CA/TG]$_n$ microsatellites

The [CA/TG]$_n$-enriched small-insert liver cDNA library was subjected to one round of screening with a [γ^{32}P]ATP-labeled [CA]$_{15}$ oligonucleotide probe. A total of 4,000 colonies from both the unenriched control and the enriched cDNA libraries were screened. Positives were counted. The calculated percentage of positive clones in the library is shown in Table 6.1.

Table 6.1 Comparison of percentages of positive clones in the [CA/TG]$_n$-enriched and unenriched cDNA libraries

DNA Library	**Positive Clone Content**
Enriched	3.5 (%)
Unenriched	0.05 (%)
Enrichment	70 (fold)

A 70-fold enrichment for the [CA/TG]$_n$ microsatellites has been achieved using the DNA affinity hybridisation selection procedure developed for genomic sequences.

(2) Size range of the inserts

After screening, positive clones were picked and propagated in LB broth. Plasmid DNA was prepared using the Qiawell-18 Plasmid Purification Kit (M&M 2.2.9). The purity, quantity and insert size of the clone were checked by digestion with restriction enzyme, *EcoRI* and electrophoresis on 1% agarose gel.

About 5% of the positive clones contained large inserts (>900 bp) or could not be digested by the enzyme. Of the remaining 95%, the sizes of inserts varied from 500-800 bp with an average of 650 bp.

(3) Characterisation of nucleotide sequences

Nucleotide sequences of individual clones were generated using the Taq cycle sequencing system on a 373 automated sequencer (ABI). Double-stranded Plasmid DNA was sequenced from both strands using T3 and T7 primers (M&M 2.2.18). Examples of DNA sequence of positive clones were shown in Figure 6.2.

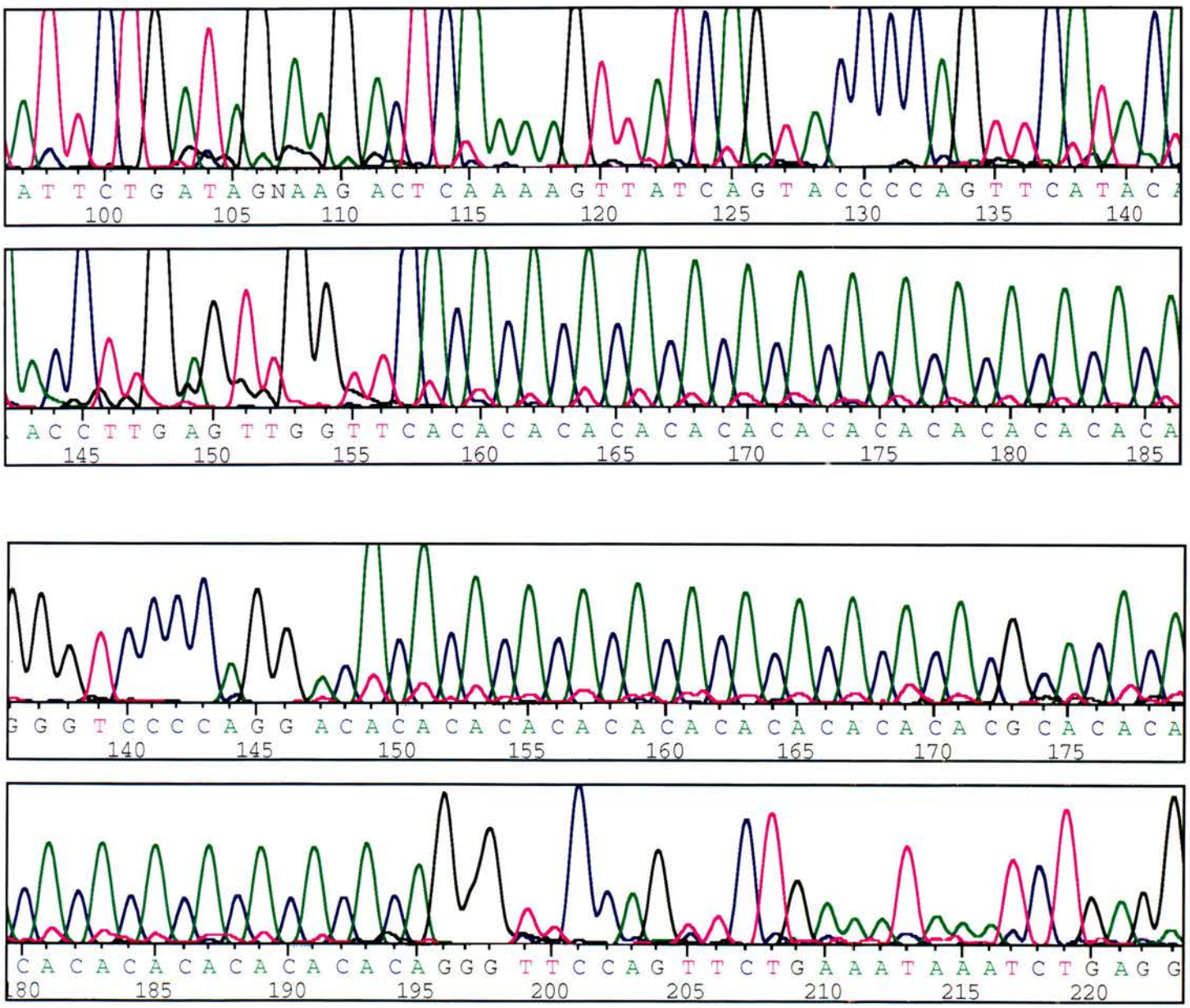


Figure 6.2 Examples of DNA sequence of microsatellite clones from the $[CA/TG]_n$-enriched cDNA library. (1) Perfect $[CA/TG]_n$ repeat; (2) imperfect $[CA/TG]_n$ repeat.

A total of 40 positive clones from the enriched cDNA library were characterised by DNA sequencing. The nucleotide sequence composition of the repeat motif from each contig is shown in Table 6.2. The positive clones sequenced fell into 15 unique contigs. 4 represent simple [CA] repeats (26%), whereas the remaining contigs represent complex repeats (74%). Less than 20% of the clones sequenced were unique clones. A number of clones were highly-represented in the enriched cDNA library. In contrast to the microsatellite clones from the $[CA/TG]_n$-enriched genomic DNA where nearly half of the clones characterised were found to be [CA/TG]-rich sequences, the results in Table 6.2 show that less than 13% of the positive clones in the cDNA library belong to this [CA/TG]-rich class.

Table 6.2 Sequence characterisation of repeat motifs of the positive clones in the $[CA/TG]_n$-enriched cDNA Library.

Contig	Number of Clones	DNA Sequence of the Repeat Motif
CA001	1	$[CA]_{19}CG[CA]_2$
CA002	1	$[CA]_{12}T[AC]_6$
CA003	2	$T_4[TG]_{14}T_3$
CA004	6	$T_4A[TG]_{24}T_4$
CA005	1	$[TG]_{19}$; $[TG]_{27}$
CA012	4	$C_4AC[CA]_{12}$
CA013	2	$[GT]_{17}$
CA016	4	$[CA]_{29}$
CA017	1	$[GT]_{13}$
CA018	2	$[TG]_{11}C[GT]_{13}$
CA023	2	$[GT]_{14}$
CA024	1	$[TG]_2[TA]_3C[GT]_{21}$
CA025	1	$[TG]_5C[GT]_9$
CA034	1	$[GT]_{10}GAA[TG]_2AA[GT]_4$
CA040	3	$[CA]_{18}[CCA]_2$
	5	[CA/TG]-rich
	5	Anonymous

(4) Size distribution of $[CA/TG]_n$ repeats

The observed size distribution of the $[CA/TG]_n$ motifs in the $[CA/TG]_n$-enriched cDNA library is shown in Figure 6.3. A typical 'normal distribution' pattern was observed for the $[CA/TG]_n$ motifs in the cDNA library. The number of repeats varied from 5 to 30 with $[CA/TG]_{11\text{-}15}$ being the most frequent category. This distribution resembles the one observed by Crooijmans *et al.* (1993) in an unenriched lambda *Zap II* phage genomic library. It is also similar to the one observed in our $[CA/TG]_n$-enriched genomic library.

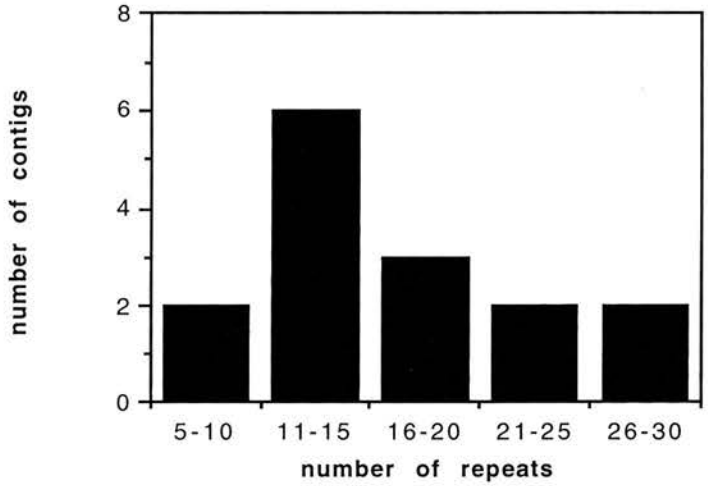


Figure 6.3 Size distribution of $[CA]_n$ motifs of positive clones in the $[CA/TG]_n$-enriched cDNA library. The number of uninterrupted repeats was plotted against the number of clones.

6.4 Characterisation of the $[CAG]_n$-enriched cDNA Library

(1) Enrichment of $[CAG]_n$ microsatellites

The $[CAG]_n$-enriched liver cDNA library was screened with a $[\gamma^{32}P]$ATP labeled $[CAG]_{10}$ oligonucleotide probe. A total of 2,000 colonies from both the unenriched

control and the enriched cDNA libraries were screened. Positive clones were counted and the calculated percentage of positive clones in the libraries is shown in Table 6.3.

Table 6.3 Comparison of percentages of positive clones in the $[CAG]_n$-enriched and unenriched cDNA libraries

DNA library	**Positive Clone Content**
Enriched	2 (%)
Unenriched	0 (%)

(2) Size range of inserts

Plasmid DNA from positive clone was prepared and restriction enzyme digested as for the positive clones from the $[CA/TG]_n$-enriched cDNA library. About 3% of the positive clones contained large inserts (> 900 bp). Of the remaining 97%, the size of inserts varied from 500 to 800 bp with an average of 650 bp.

(3) Sequence characterisation of positive clones

Nucleotide sequences of individual positive clones were generated using the Taq cycle sequencing system on a 373 automated sequencer (ABI). Double-stranded plasmid DNA was sequenced from both strands using T3 and T7 primers (M&M 2.2.18). An example of nucleotide sequence is shown in Figure 6.4.

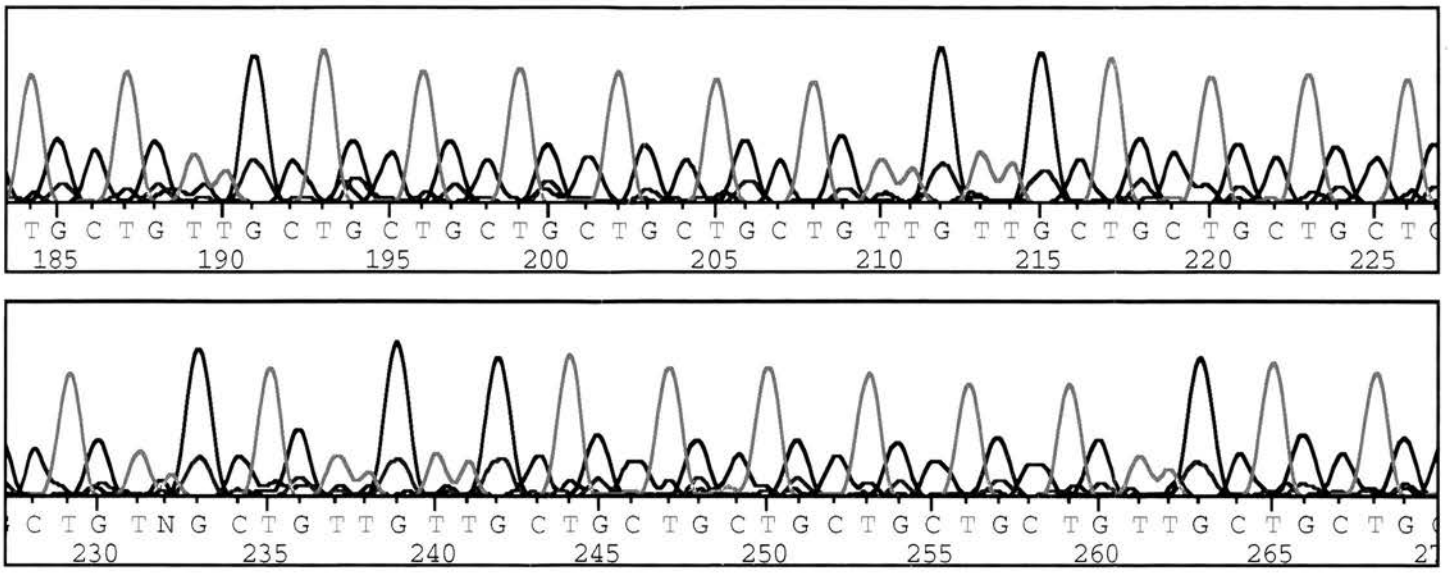


Figure 6.4 Example of DNA sequence of a positive clone from the $[CAG]_n$-enriched cDNA library.

A total of 17 positive clones from the $[CAG]_n$-enriched cDNA library were sequenced. The nucleotide sequence composition of the trinucleotide repeat motifs is shown in Table 6.4.

Table 6.4 Sequence characterisation of repeat motifs of positive clones from the $[CAG]_n$-enriched cDNA library. (X= G or A)

Contig	Number of Clones	DNA Sequence of the Repeat Motif
CAG001	2	$[CAX]_{43}$
CAG007	1	$[CAX]_{37}$
CAG009	1	$[CAX]_8ATGCTGAGG[CAX]_{23}$
CAG015	4	$[CAX]_9AGT[CAX]_2AAC[CAX]_7$
CAG018	4	[CAX]-rich
CAG019	4	[CAX]-rich
CAG020	1	$[CAX]_{17}CGG[CAX]_{20}$

As shown in Table 6.4, 7 unique contigs were formed. Among the 7 unique contigs, 3 were represented by single clones whereas the other 4 contigs were represented by multiple clones. This reflects a very low complexity of the $[CAG]_n$-enriched cDNA library.

Both contigs CAG001 and CAG007 contain a simple [CAX] (X=A or G) repeat with a repeat number of 43 and 37 respectively, whereas CAG20 contains 37 [CAX] repeats with a single interruption by a [CGG] triplet in the middle. The rest of the contigs are either [CAX]-rich sequences or contains complex [CAX] repeats.

6.5 Database searches for sequence homologies with the Microsatellite-containing cDNA Sequences

The BLAST (Basic Local Alignment Search Tool) e-mail server at the National Centre of Biotechnology Information (NCBI), National Library of Medicine, NIH in Bethesda was used for sequence similarity searches (Altschul *et al.*, 1990). BLASTN uses nucleotide sequence as query sequence, while BLASTX translates the nucleotide sequence in six reading frames and then uses the translated protein sequence to perform a sequence similarity search. The non-redundant (nr) peptide sequence databases searched include: Pdb, Swissprot, Pir, Spupdate, Genpept, Gpupdate, Tfd, Acr and Aalu. The nonrandom (nr) nucleotide sequence databases include: Pdb, Genbank, Gbupdate, Embl, Emblu, Vector, Repbase, Katatnue, Epd and Dbest. In addition, sequence similarity searches were carried out by running the FASTA program of GCG Package (University of Wisconsin Genetics Computing Group).

To perform a database search, sequence information of cDNA microsatellite clones was collected from a 373 DNA sequencer, and edited using Seqman software (DNA Star). The consensus sequence of each individual contig was submitted in the

requested format to BLAST or FASTA. A summary of the database searches and the general results from the database searches are given in Table 6.5 and Table 6.6.

Table 6.5 Summary and comparison of the results of database searches for $[CA]_n$- and $[CAG]_n$-containing cDNA sequences

	$[CA]_n$-containing cDNA Sequences	**$[CAG]_n$-containing cDNA Sequences**
Known gene	4 (16%)	1 (11%)
Related sequence	4 (16%)	1 (11%)
Unknown sequence	17 (68%)	7 (78%)
Total	25	9

Table 6.6 General results of sequence similarity search for chicken liver expressed sequences

Query Contig	Length (bp)	Putative Identification or Homology & Database Entry	Identified Species	Program & Matrix	HSP Score	P Value
CA001	486	Uknown		BLAST	ns	ns
CA002	464	Unknown		BLAST	ns	ns
CA003	524	FMR1 gene product: GP:L19493:HUMFMR1R_1 (known)	HUMAN	BLASTX;BLOSUM62	91	1.10E-19
CA004	386	Pregnancy-specific beta-1-glycoprotein:PSG1:GP:M22312:HUMPBG2_2 (related)	HUMAN	BLASTX;BLOSUM62	74	3.60E-03
CA005	414	NBL4 mRNA;BAND 4.1 Superfamily:GB:D28818:MUSNBL4 (known)	MOUSE	BLASTN	150	1.10E-02
CA008	408	Unknow		BLAST	ns	ns
CA011	318	NBL4 mRNA;BAND 4.1 Superfamily:GB:D28818:MUSNBL4(known)	MOUSE	BLASTN	641	2.10E-44
CA012	567	Unknown		BLAST	ns	ns
CA013	295	Unknown		BLAST	ns	ns
CA014	253	Unknown		BLAST	ns	ns
CA015	180	Unknown		BLAST	ns	ns
CA016	545	Unknown		BLAST	ns	ns
CA017	641	ATP13 protein precursor:SP:P22136:ATPU_YEAST (related)	YEAST	BLASTX;PAM120	93	3.50E-06
CA018	545	Thrombin:PDB:1BBR:E (related)	BOVINE	BLASTX;BLOSUM62	148	1.50E-03
CA019	356	Unknown		BLAST	ns	ns
CA020	217	Unknown		BLAST	ns	ns
CA023	509	PLT gene product:GP:M92449:HUMPLT_1	HUMAN	BLASTX;BLOSUM62	90	3.10E-12
CA024	259	Unknown		BLAST	ns	ns
CA025	636	Unknown		BLAST	ns	ns
CA028	380	Unknown		BLAST	ns	ns
CA033	292	Unknown		BLAST	ns	ns
CA034	414	Dimethylanine mono-oxygenase 3:SP:P31512:FM03 HUMAN (related)	HUMAN	BLASTX;PAM250	50	3.20E-03
CA035	356	Unknown		BLAST	ns	ns
CA040	323	Unknown		BLAST	ns	ns
CA042	480	Unknown		BLAST	ns	ns
CAG001	600	Transcriptional factor MEF2X:GP:L16794:HUMANMEF2X_1 (known)	HUMAN	BLASTX;BLOSUM62	330	3.50E-39
CAG007	529	Unknown		BLAST	ns	ns
CAG008	404	Unknown		BLAST	ns	ns
CAG009	141	Unknown		BLAST	ns	ns
CAG015	502	Transcription factor NTRC,Nitrogen regulation protein:GNL; TFD:P00079:SP:P06713:NTRC_E.coli (related)	E.COLI	BLASTX;BLOSUM62	63	4.7E-03
CAG017	138	Unknown		BLAST	ns	ns
CAG018	574	Unknown		BLAST	ns	ns
CAG019	522	Unknown		BLAST	ns	ns
CAG020	536	Unknown		BLAST	ns	ns

HSP: high-scoring segment pair. P: probability. Known genes: HSP=90-641, P value=4.6×10^{-58} to 7.9×10^{-8}; Related genes: HSP=50-173, P value=3.5×10^{-6} to 3.6×10^{-3}. ns: not significant.

6.6 Sequence Homologies of cDNA Clones

6.6.1 The Chicken Homologue of the Human MEF2D Transcription Factor

Pairwise sequence comparison was carried out between hMEF2D and the consensus sequence of CAG001. The sequence alignment is given in Figure 6.5. Multiple sequence comparisons were carried out for hMEF2A, hMEF2B, hMEF2C, hMEF2D, mMEF2C, SL-1, SL-2 and CAG001. The results of this multiple sequence comparison were used in the construction of a phylogenetic tree (Figure 6.6). The cDNA sequence was translated into an amino acid sequence and the result of amino acid sequence comparison between chicken, *Xenopus* and human is shown in Figure 6.7.

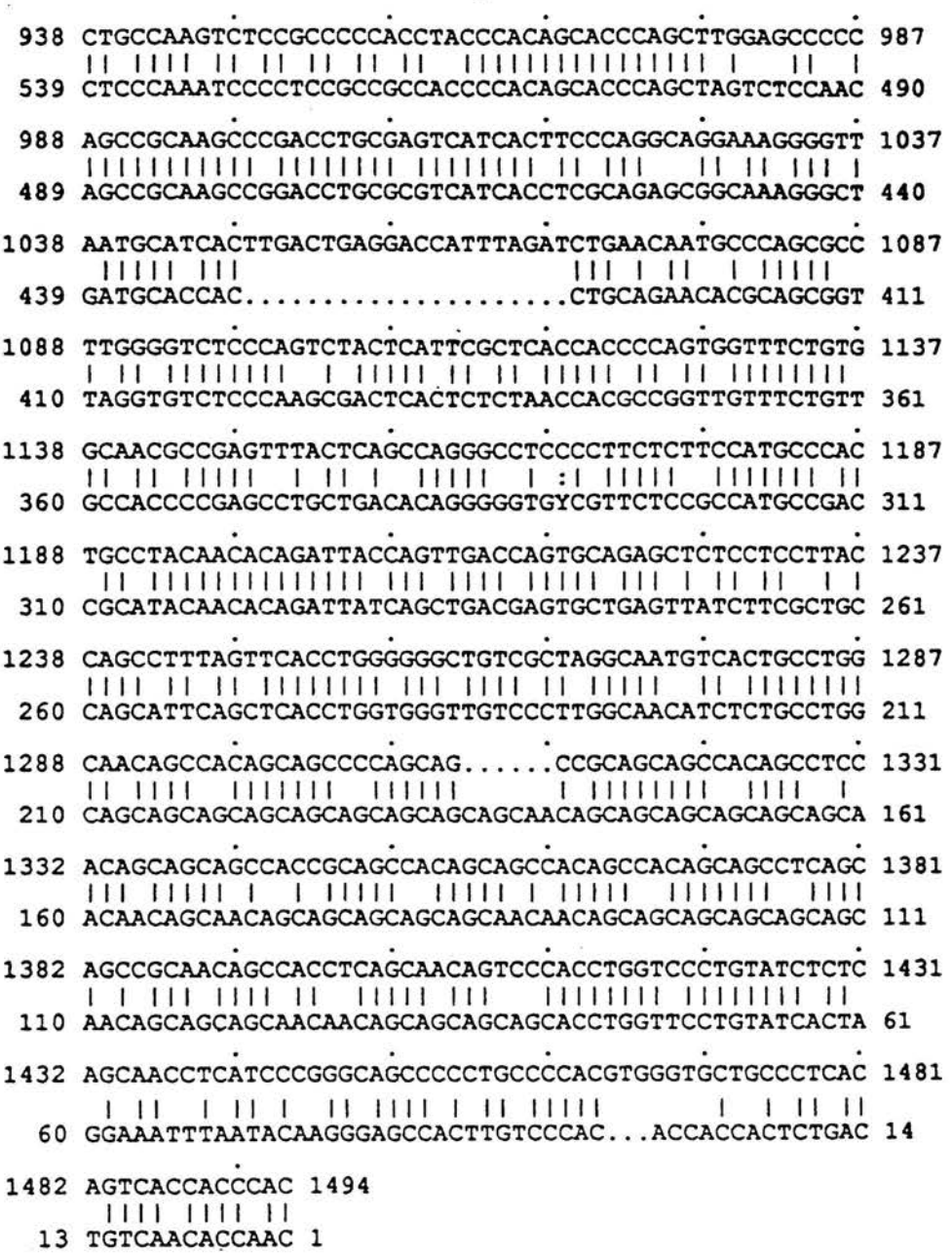


Figure 6.5 Comparison of nucleotide sequences of CAG001 (lower) and human hMEF2D (upper). DNA sequence alignments were performed using the BESTFIT program of GCG package.

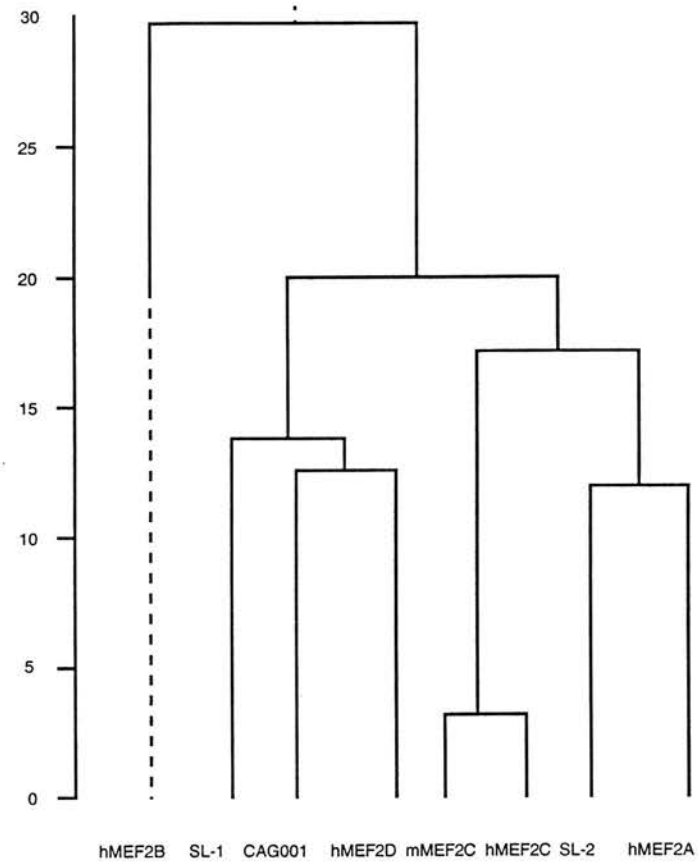


Figure 6.6 Phylogenetic tree of the MEF2 gene family based on the nucleotide sequence similarity. Multiple-sequence alignment was carried out using the CLUSTAL method of Seqman (DNA Star).

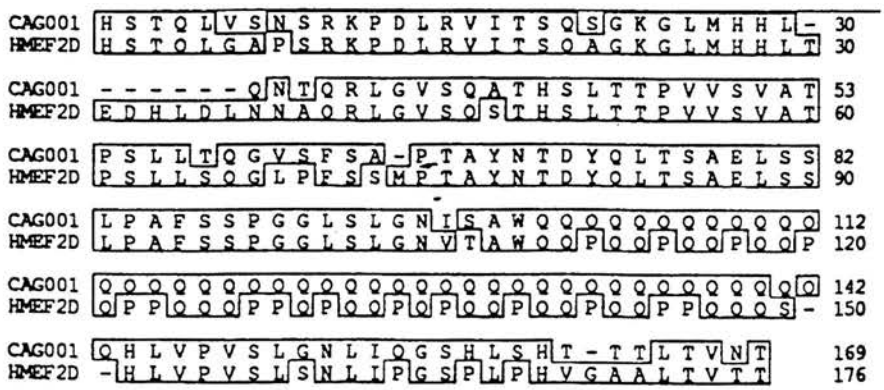


Figure 6.7 Comparison of amino acid sequences of the poly-glutamine/proline tract and its flanking regions in the human, *Xenopus* and chicken MEF2D peptides ('Clustal' Method, Seqman). Sequence identity is boxed

Figure 6.5 shows that contig CAG001 has 73% sequence similarity with the coding region of human MEF2D gene over a length of 563 bp. Within this region a [CAG] trinucleotide repeat motif is conserved in both the human and chicken gene. Three deletions/insertions have taken place within the aligned region. One of the insertion/deletion events occurred within the poly[CAX] repeat motif. Figure 6.6 shows that human MEF2D (hMEF2D) and its *Xenopus* homologue (SL-1) are the most closely-related genes to contig CAG001 within the MEF2 gene family. Together these results support the conclusion that contig CAG001 is the chicken homologue of human MEF2D.

The sequence homology of hMEF2D with CAG001 is located within nucleotide 938 and 1759 as shown in Figure 6.8.

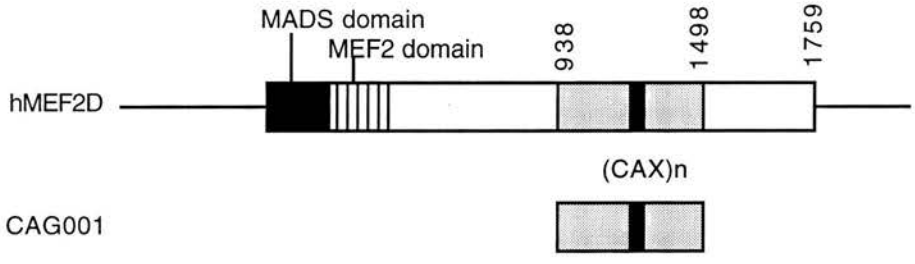


Figure 6.8 Location of the nucleotide sequence alignment of CAG001 and hMEF2D. Line indicates 5'/3' untranslated region; stippled rectangle indicates sequence homology.

6.6.2 The Chicken Homologue of the Human FMR1

The nucleotide sequence of the human fragile X mental retardation gene (hFMR1) was retrieved from the EMBL database. Pairwise DNA sequence alignment was

carried out between hFMR1 and AC003. The sequence alignment is shown in Figure 6.9.

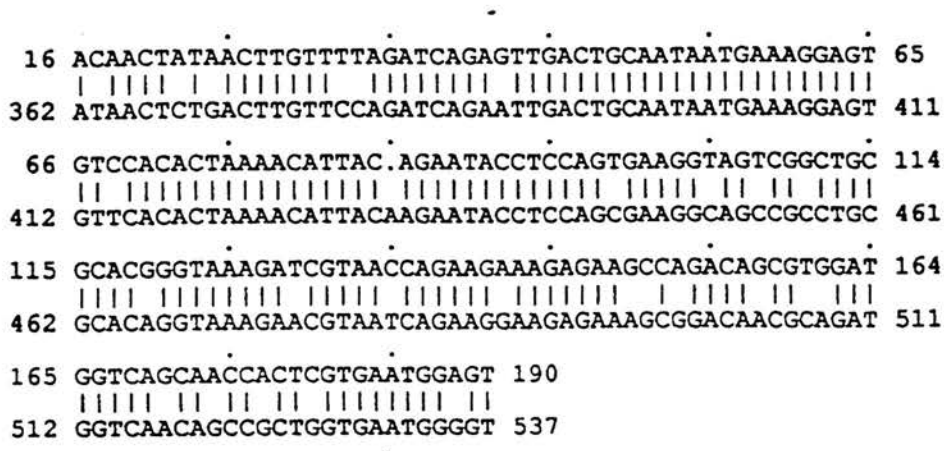


Figure 6.9 Comparison of nucleotide sequence of human FMR1 (upper) cDNA and the consensus sequence of contig CA003 (lower). (see Figure 6.5 for method description).

Contig CA003 has a sequence similarity of 86% over a length of 176 bp (nucleotides 362 to 537) with the 3' region of the human fragile X syndrome gene. The sequence alignment lies 3' to the putative RGG domain of the hFMR1 gene. However, the sequence region between nucleotides 1 and 362 showed poor homology with the human gene.

A previous study has shown that the human FMR1 gene consists of 17 exons spanning 38 kb DNA (Eichler *et al.*, 1993). As shown in Figure 6.10, the sequence homology between our partial cDNA clone CA003 and the human FMR1 gene is located at the 3' end of the coding region. It covers the whole region of the last exon (exon XVII) and part of the last intron (intron XVI). Within this homologous region, an alternative RNA splicing site is found. This is conserved between the human and the chicken genes.

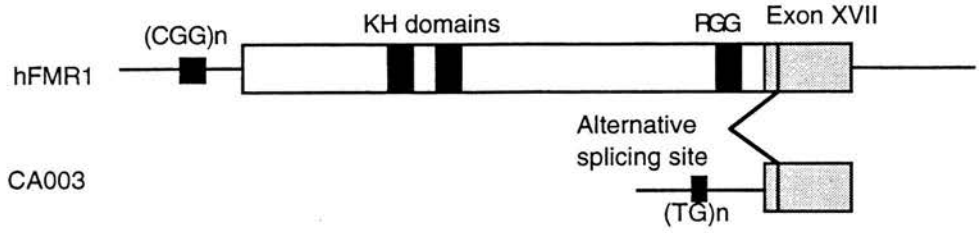


Figure 6.10 Location of the sequence homology between CA003 and hFMR1. Rectangle indicates coding region, line indicates intron or 5'/3' untranslated regions.

So far, homologous FMR1 genes have only been cloned in human and mouse. Figure 6.11 shows the nucleotide sequence alignment for the last exon (exon XVII) of chicken, mouse and human FMR1 homologous genes. As shown, a high degree of sequence conservation is observed.

```
AC003 AATACCTCCAGCGAAGGCAGCCGCCTGCGC  30
HFMR1 AATACCTCCAGTGAAGGTAGTCGGCTGCGC  30
MFMR1 AGTGCCTCCAGTGAAGGGAGCCGGCTGCGC  30

AC003 ACAGGTAAAGAACGTAATCAGAAGGAAGAG  60
HFMR1 ACGGGTAAAGATCGTAACCAGAAGAAAGAG  60
MFMR1 ACGGGTAAAGATCGTAACCAGAAGAAGGAA  60

AC003 AAAGCGGACAACGCAGATGGTCAACAGCCG  90
HFMR1 AAGCCAGACAGCGTGGATGGTCAGCAACCA  90
MFMR1 AAGCCAGACAGCGTAGATGGGCTGCAACCG  90

AC003 CTGGTGAATGGGGTGAG-TAG            110
HFMR1 CTCGTGAATGGAGTACCCTAA           111
MFMR1 CTGGTGAATGGAGTACCCTAA           111
```

Figure 6.11 Comparison of the nucleotide sequences of exon XVII of chicken, mouse and human FMR1 ('Clustal' method, Seqman). Sequence identity is boxed.

6.6.3 The Chicken Homologue of the Mouse NBL4 Gene

The cDNA sequence of mouse NBL4 (MUSNBL4), a member of the "Band 4.1" superfamily, was retrieved from the EMBL database. Pairwise sequence comparison was carried out between MUSNBL4 and CA011. The sequence alignment is shown in Figure 6.12. CA011 and MUSNBL4 have an overall sequence similarity of 73% over a 377 bp DNA region.

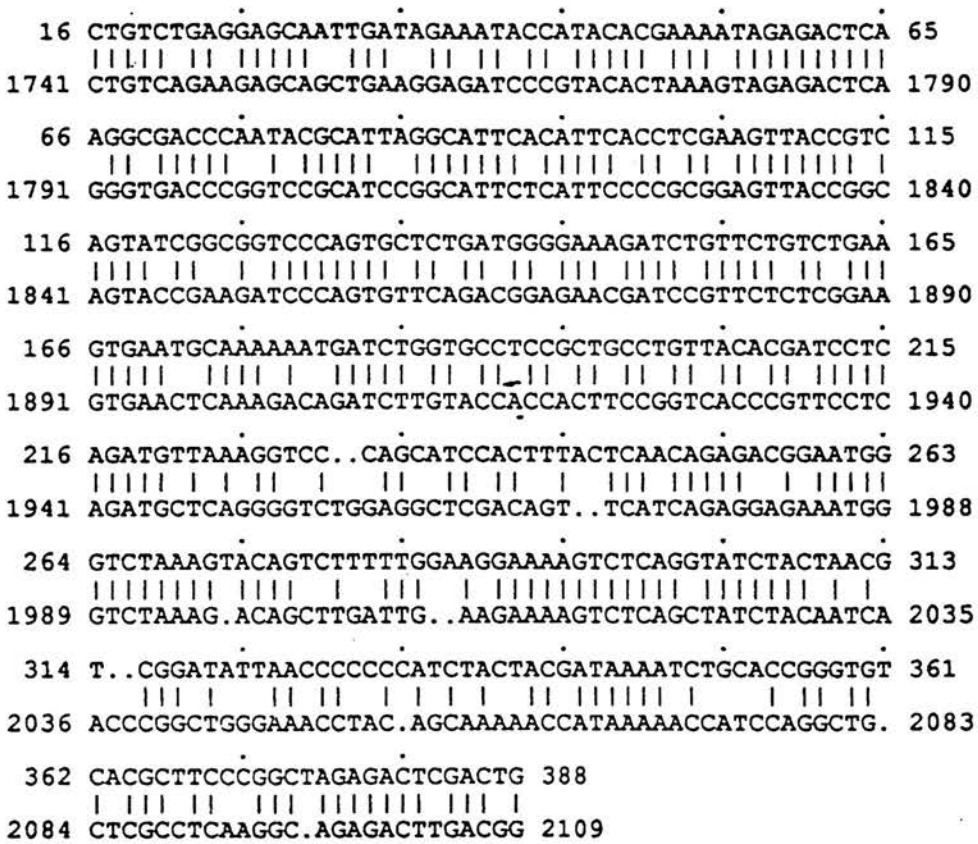


Figure 6.12 Comparison of nucleotide sequence of mouse NBL4 (lower) cDNA and the consensus sequences of CA011 (upper). (see Figure 6.5 for method description).

The open reading frame (ORF) finding program of the SEQMAN package (DNA Star) was used to define the open reading frame for MMNBL4. As shown in Figure

6.13, CA011 has high sequence homology with MMNBL4 in the 3' untranslated region.

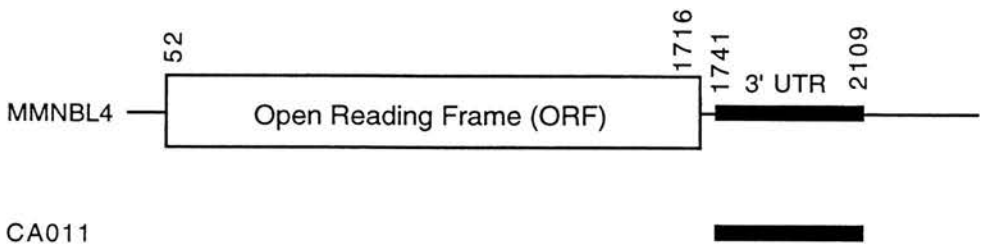


Figure 6.13 Putative coding frame of MMNBL4 and the location of the sequence homology between CA011 and MMNBL4.

6.6.4 The Chicken Homologue of the Human PLT Gene

The nucleotide sequence of human PLT gene cDNA was retrieved from EMBL database. Pairwise sequence comparison was carried out between hPLT and CA023. The sequence alignment was shown in Figure 6.14.

```
211 TGGTTTCGTGTGGAGACAAACTATGACCATTGGACTACCCCTCCTCCCTT 260
    ||||| || || |||||||| || ||||| ||||   |  | || |
601 TGGTTCCGAGTTGAGACAAATTACGACCACTGGAAGCCAGCACC.CAAGG 649

261 GTGATGACCGAAGAACGCCACCCATGAAAGCTCTCAATGCTACTGGACAG 310
      |||||||| |||||  |  |||| || || || |||||||| |||||
650 AAGATGACCGGAGAACATCTGCCATCAAGGCCCTTAATGCTACAGGACAA 699

311 CAGAACATCAATTTTGATACACTTTTTCAG 340
       ||| |||   | ||  ||||||| |||
700 GCAAACCTCAGCCTGGAGGCACTTTTCCAG 729
```

Figure 6.14 Comparison of nucleotide sequence of human PLT cDNA (lower) and the consensus sequence of contig AC023 (upper). (see Figure 6.5 for method description).

Contig CA023 has an overall 68% sequence similarity over a 130 bp long region (nucleotides 211 to 340) with the 3' coding region of the human PLT gene. The sequence region between nucleotides 1 and residue 211 did not show significant homology to the corresponding region of the human gene.

As shown in Figure 6.15, the nucleotide sequence homology between hPLT and CA023 is located at the 3' end of the coding region of hPLT. Within this homologous region, two N-glycosylation sites are found in the human PLT gene, one of which is conserved in the chicken PLT gene.

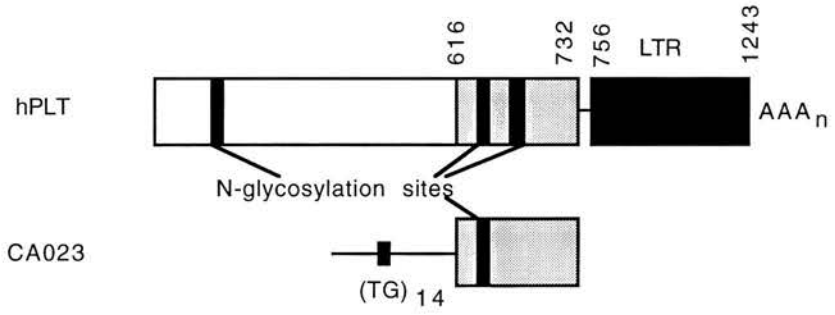


Figure 6.15 Location of the sequence homology between hPLT and CA023

6.7 Putative Annotation of Unknown cDNA Sequences

BLASTX searching did not find any sequence homology in the database for CAG007, CAG009, CAG015, CAG018, CAG019, CAG020, CA013, CA040 and CA042. The nucleotide sequences of these microsatellite-containing cDNA clones were subjected to analysis of potential exons using GRAIL1 (Gene Recognition and Analysis Internet Link) program (Uberbacher and Mural, 1991). The results are shown in Table 6.7.

Table 6.7 Potential Exons based on GRAIL1 analysis of $[CAX]_n$- or $[CA]_n$-containing cDNA sequences.

Sequence	Position	Strand	Strand Probability	Frame	Quality	ORF
CAG007	51-277	R	0.63	1	Excellent	1-277
CAG009	51-91	R	0.53	1	Good	1-124
CAG015	331-461	R	0.57	2	Excellent	314-518
CAG018	51-141	R	0.53	2	Excellent	1-141
CAG019	51-161	F	0.63	3	Excellent	1-165
	51-101	R	0.66	2	Good	11-137
CAG020	311-361	R	0.86	1	Good	304-367
	56-281	R	0.64	2	Excellent	56-281
CA013	61-141	R	0.83	1	Excellent	1-295
CA040	51-71	F	0.77	3	Marginal	1-138
CA042	261-341	F	0.99	2	Excellent	179-392

ORF: open reading frame; R: reverse; F: forward.

A total of eighteen $[CA]_n$-containing and six $[CAX]_n$-containing cDNA sequences were analysed. As shown in Table 6.7, all $[CAX]_n$-containing sequences, however, only three $[CA]_n$-containing sequences (17%, CA013, CA040 and CA042) were found to have high coding potentials.

Chapter 7

Studies of Microsatellite Polymorphisms

We have identified 40 unique microsatellite sequences from genomic DNA and 22 from cDNA. The degree of polymorphisms of these microsatellites needs to be investigated by polymerase chain reaction assays.

7.1 Designing Primers for Polymerase Chain Reaction for Microsatellite Assays

Sequence data of suitable microsatellite clones from both genomic and cDNA libraries were first checked by eye to ensure that enough flanking sequence existed for designing PCR primers. Then the PRIMER program was used to select unique PCR primers flanking each microsatellite. Attempts were made to select PCR products that were 100 to 270 bp in length. The selected primers and their expected product lengths are listed in Table 7.1.

7.2 Polymerase Chain Reactions and Product Analysis

Primers were synthesised, labeled and HPLC purified using standard methods (Oswel DNA, Edinburgh UK). The polymerase chain reaction conditions used were 50 mM KCI, 10 mM Tris-HCI (pH 9.0), 0.1% Triton X-100, 1.5 mM $MgCI_2$, 200 µM each dNTP, 10 pmol of each primer, 0.5 U of Taq polymerase, 50-100 ng of genomic DNA in a final volume of 10 µl. Reactions were initially denatured for 1 minute at 94^0C, then 23 cycles of 94^0C for 30 seconds; 60^0C for 1 minute and 72^0C for 3 minutes, with a final 10 minute elongation step at 72^0C in a thermal cycler (TRIO-BLOCK). The samples were heated to 94^0C for 2 minutes, 1 µl was loaded onto a 6% sequencing gel on an automated sequencer (Applied Biosystem). Data was collected and analysed using GENESCAN software. Both the male and female parents in the East Lansing and Compton crosses were genotyped using microsatellite, ROS001. The detection of polymorphisms for ROS001 is shown in Figure 7.1.

Table 7.1 Polymerase chain reaction primers for $[CA/TG]_n$ microsatellites

Locus	Label	Sequence of Repeat	Primers	Product Size (bp)
ROS001	M043-5F	$(AT)_5AA(AT)_4$-	TGAAAGGGATGGAGTATGTCA	174
	M043-3	$(AC)_{19}$	CCTTGTGATCTCTCCACACTTC	
ROS002	M044-5J	$(AC)_{19}$	TGGAAGCTACTGACAGTTTTCC	254
	M044-3		GCATCCTTTGTGTCCTGACA	
ROS003	M046-5F	$(AC)_{13}$	GCAAAGTTATTCAGGAACTTGC	250
	M046-3		AAGTGGTCCCCTGATTTAACA	
ROS004	M049-5J	$(CA)_{21}$	ACTTAAGCCGGCACTTGTGT	261
	M049-3		ACACCTTTCCTCTGGTCCCT	
ROS005	M050-5J	$(AC)_{18}$	GAAGTGTGGGGTTTGCTGTT	188
	M050-3		TTGGATCAGATGCATCCAAA	
ROS006	M116-5F	$(AC)_{24}$	ACACCCATGTGTCAACTGGA	161
	M116-3		AACAGAGGCACCCAAAAATG	
ROS007	M120-5F	$(CA)_{28}$	GCTCTCCTAGGGCACAGATG	257
	M120-3		AAACCCGCTGTGTGTGTGT	
ROS008	M121-5J	$(CA)_{17}$	GGACAATTCCCCAGTCACAC	101
	M121-3		GCATATTTGATTGTGAAATGGG	
ROS009	M162-5F	$(CA)_{17}$	AACCCAATTGTCCATTATCAGG	268
	M162-3		ACAGGGAGACAGAGAGATGCC	
ROS010	M164-5F	$(AC)_5AG(AC)_{27}$	ATTGCTACTCACCACCACCC	226
	M164-3		TCAATAAGCACCCATGTAGTGC	
ROS011	M167-5J	$(AC)_{14}AGACG$-	AAACTGAAAATGAATGCGGC	173
	M167-3	$(AC)_{16}$	AAAAGCGTGTGATTTGTCCC	
ROS012	M172-5J	$(AC)_{20}$	TATGGAGTACAAACTCCCATCC	250
	M172-3		TACTCAGGCGTTGAACACAG	
ROS013E	CA001-5F	$(CA)_{18}$	CAGCCTTGAAAGGGTCAAAG	116
	CA001-3		AATCAGGGTGTGCGTGTGT	
ROS014E	CA003-5J	$(AC)_{14}$	AAGCCCTTCGACTACTTGCA	222
	CA003-3		AGACAACTAGCTGAAACAGGC	

(To be continued)

(Continued from page 91)

ROS015E	CA004-5T	$(CA)_{24}$	AGGTAGTTCTTCAGGGGAAAGA	250
	CA004-3		TGTTTCAGGTAAGCACAGACAA	
ROS016E	CA005-5J	$(AC)_{22}..(CA)_{19}$	CAAGCAGGCTGTGTTTCTGA	246
	CA005-3		GAAGCGAGAAAAACTGTGCC	
ROS017	M176-5F	$(AC)_5AG(AC)_6AT-$	CCAGGTTGGACACTCTAGATCC	197
	M176-3	$(AC)_{18}$	TAGTGAAGTTGTCAGAGCCTGC	
ROS018	M183-5J	$(CA)_{11}CTCACAC-$	GCAACTGTCACTGCAAGATCA	125
	M183-3	$ACT(CA)_4$	AGGGGTTTCTGAGGGAGTGT	

Note: ROS001-012, 017 and 018 are microsatellites from genomic DNA, whereas ROS013-017E are microsatellites from cDNA. Fluorescent label: J(OE), green; F(AM), blue; T(AMRA), yellow.

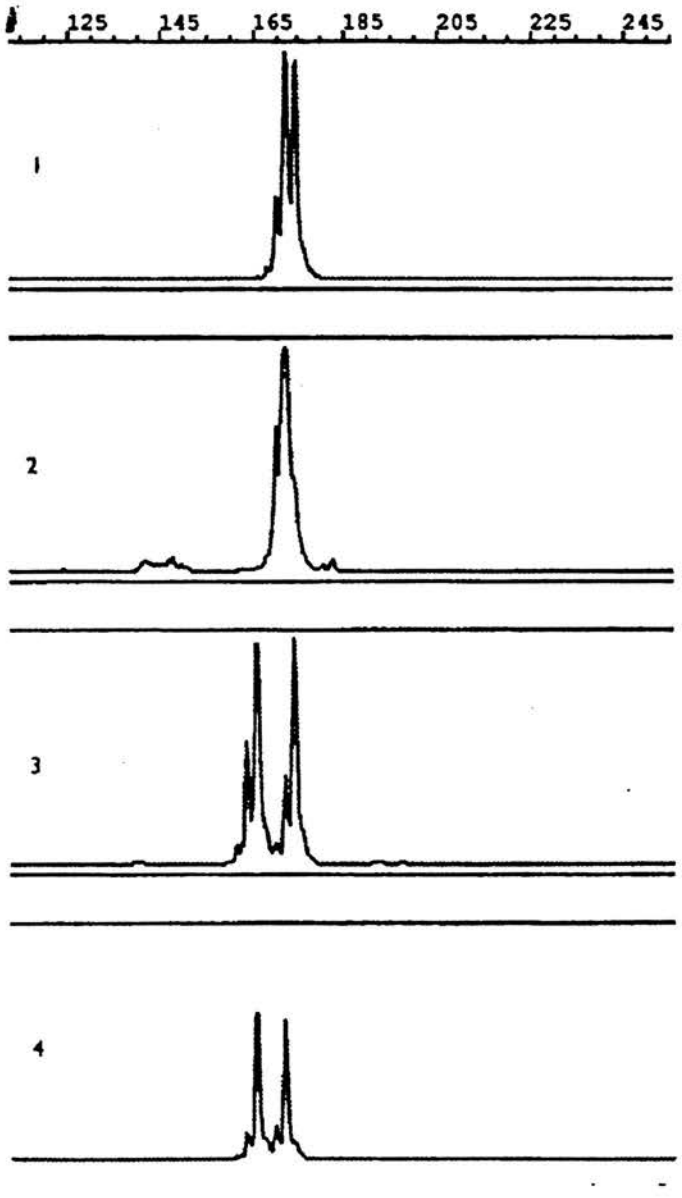


Figure 7.1 PCR analysis of microsatellite polymorphisms. Lane 1 and 2: East Lansing male and female parent; Lane 3 and 4: Compton male and female parent. The product size (bp) is indicated at the top.

As shown in Table 7.2, three alleles (A_1=172 bp, A_2=174 bp and A_3=168 bp) have been detected in this experiment. In the East Lansing cross, the male parent is a heterozygote (A_1A_2) whilst the female parent is a homozygous (A_1A_1). In the Compton cross, both the male and female parents are heterozygous (A_2A_3, A_1A_3).

Table 7.2 Genotypes of the male and female parents of East Lansing and Compton Cross.

Gel Lane	Identity	Genotype	Size of allele
1	Male (E)	A_1A_2	172, 174
2	Female (E)	A_1A_1	172
3	Male (C)	A_2A_3	168, 174
4	Female (C)	A_1A_3	168, 172

E: East Lansing cross; C: Compton cross.

The microsatellite, ROS001, was assigned into linkage group E2 in the East-Lansing cross by linkage analysis. The relative position of ROS001 in the linkage group is shown in Figure 7.2.

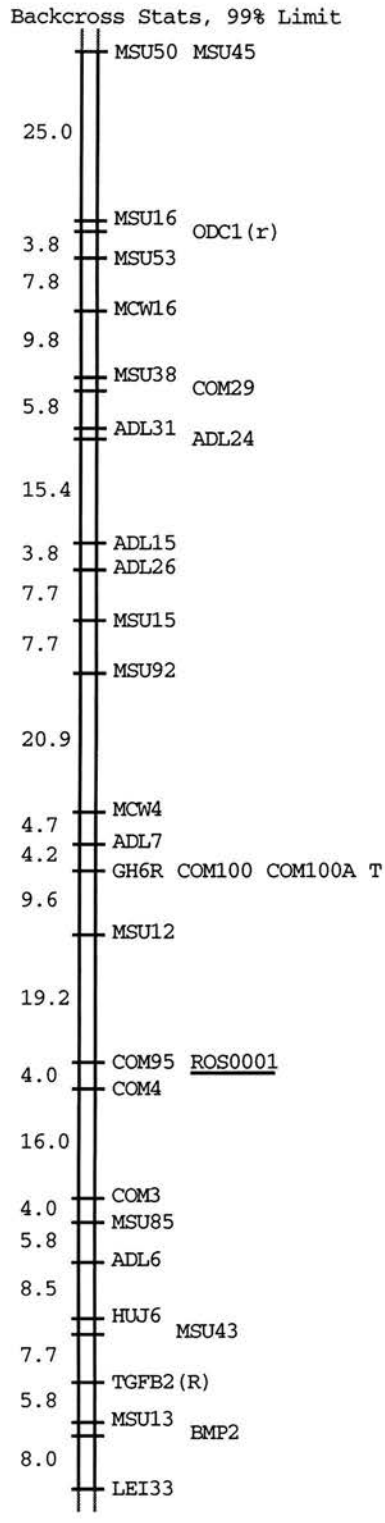


Figure 7.2 Map of the linkage group E2 in the East-lansing mapping cross. ROS001 is underlined.

Chapter 8

Studies of Differential Expression of Chicken Myocyte Enhancer Factor 2D (MEF2D)

Twenty two unique microsatellite-containing cDNA clones have been isolated. . Searches of DNA databases identified four sequence homologies. The remaining cDNA sequences have no identified homologues. First, these cDNAs will be mapped onto the chicken genetic linkage map as expressed sequence tags (ESTs) by using microsatellite markers in linkage studies. Mapped ESTs can then be used to map other genes or sequences in future studies. Second, mapped cDNA sequences can serve as good candidate genes for a particular trait or phenotype. To make use of these cDNA sequences as candidate genes, their patterns of expression need to be determined. To illustrate this process, the expression of CAG001 was studied.

8.1 Introduction

Since CAG001 is identified to be the chicken homologue of the human myocyte enhancer factor 2D (MEF2D, see Chapter 6), it is an appropriate candidate gene for the control of muscle development or growth rate in chicken. In mammals, four genes MEF2A, MEF2B, MEF2C and MEF2D belong to the myocyte enhancer factor 2 (MEF2) gene family (Martin *et al.*, 1993; Chambers *et al.*, 1992; Leifer *et al.*, 1993; Yu *et al.*, 1992; Pollock and Treisman, 1991; McDermott *et al.*, 1993). The protein products of these genes together share two highly conserved DNA-binding domains, the MADS domain and the MEF2-specific domain. These transcriptional factors are known to regulate the expression of some muscle-specific genes by binding to specific binding sites within their promoter regions.

The mammalian MEF2 genes are expressed in a restricted tissue-specific pattern, although some transcripts are present in a variety of tissues. For instance, the human MEF2A mRNA is ubiquitous with preferential accumulation in skeletal muscle, heart and brain (Yu *et al.*, 1992). In contrast to human MEF2A, the human and mouse MEF2C mRNAs are largely restricted to skeletal muscle and brain (McDermott *et al.*,

1993). Breitbart *et al.* (1993) determined the distribution patterns of human MEF2D mRNA, showing that MEF2D mRNA , like MEF2A, was ubiquitous. Taken together, the different patterns of MEF2 gene expression suggests that they might have distinct functions.

8.2 Design and Validation of Chicken MEF2D-Specific Primers for RT-PCR

Northern blot analysis lacks sufficient sensitivity to detect transcripts in small amounts of tissue, especially when those transcripts are present at a low level or have short half-life. Also, it is too labour intensive for screening large number of cDNAs. The technique of reverse transcription and polymerase chain reaction (RT-PCR), which allows the detection of rare gene transcripts, was used in this experiment to detect the presence of chicken MEF2D mRNA. Also, this will be a model for future work to rapidly detect the general pattern of expression for microsatellite-containing chicken expressed sequences, in particular, unknown genes. The $[CAX]_n$ tri-nucleotide repeat (X=G or A) within the chicken MEF2D gene 3' coding region was selected to be the target for PCR amplification. Primers were designed that bracket the [CAX]-trinucleotide repeat using the PRIMER program. Since ß-actin mRNA is believed to be present in all tissues, primers for this gene were used as a control for efficient cDNA synthesis between different RNA samples (Burt *et al.*, 1992). Furthermore, for every RNA sample, a control PCR reaction which contains all reagents but untranscribed RNA as template was included to monitor for possible contamination by genomic DNA. The primers and their expected PCR products are listed in Table 8.1.

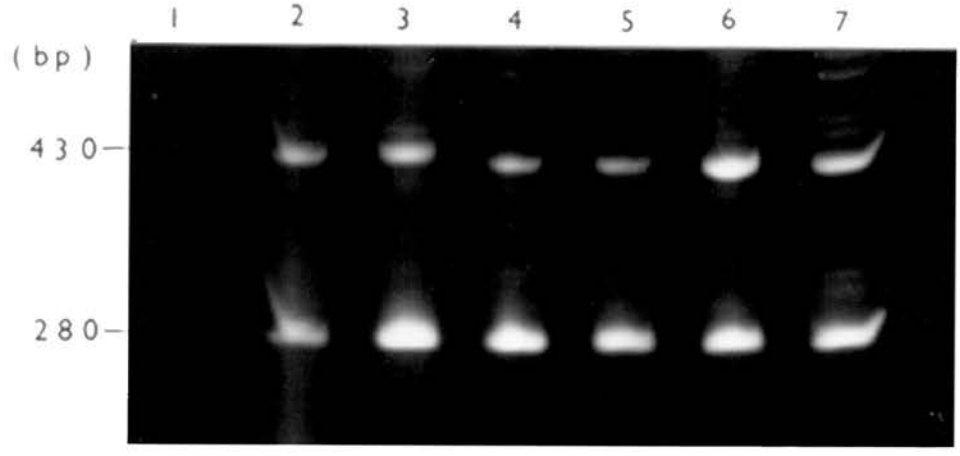
Table 8.1 Primer pairs for PCR amplification

Gene	PCR Primer	Product Size (bp)
5'-ß-actin	TGTGGTATCCATGAAACTA	280
3'-ß-actin	ATTCATCGTACTCCTGCTT	
5'-MEF2D	CAGAGTGGTGGTGTGGGAC	430
3'-MEF2D	GTCTCCCAAGCGACTCACTC	

8.3 Detection of Chicken MEF2D mRNA by RT-PCR

Total RNA was isolated from tissues using the standard guanidine thiocyanate extraction procedure (Sambrook *et al.*, 1989). First-strand cDNA synthesis was carried out using the Riboclone cDNA Synthesis Kit (Promega) (M&M 2.2.14). Five µg of total RNA was reverse transcribed by AMV-reverse transcriptase in a total volume of 25 µl primed with oligonucleotide primer, $d[T]_{25}$. For each RT reaction, a blank was prepared using all reagents except the RNA sample. PCR was carried out as described (Sambrook *et al.*, 1989) using the primer pairs shown in Table 8.1. The reactions were denatured at 94^0C (1 minute), annealed at 60^0C (2 minutes) and extended at 72^0C (3 minutes) for 30 cycles in a Biometra TRIO-BLOCK thermal cycler. A PCR blank which included all reagents and RT blank template DNA was run as a control. For each of the RNA samples, a control PCR which uses untranscribed RNA as template was set up. After RT-PCR, 10 µl of the reaction mixture from ß-actin PCR and 10 µl from MEF2D PCR of the same RNA sample were mixed together and resolved on 5% non-denaturing polyacrylamide gels along with molecular weight markers.

In all experiments, the RT-PCR reagent blank was negative. Although PCR products larger than the size expected for cDNA were detected following RT-PCR of all the RNA samples, the same-sized products were not detected from the control PCR reactions which used untranscribed RNA samples as templates, indicating that the RNA samples used in these experiment were indeed free of contaminating genomic DNA. The identity of both ß-actin and MEF2D transcripts was confirmed by the appropriate size of their PCR products (280 and 430 bp respectively).



Expression of ß-actin (lower band) and MEF2D (upper band) in tissues of chicken (1 week of age). From left to right: RT-PCR blank, breast muscle, brain, liver, lung, ovary, kidney, and heart. mRNA was reverse transcribed and amplified by PCR. Products were resolved by 5% non-denaturing acrylamide gel.

Various tissues of chicken were investigated. As shown in the above figure, transcripts of chicken MEF2D gene were detected in both muscle (breast muscle, heart) and non-muscle (brain, liver, ovary) tissues of chicken (1 week). However,

mRNA of chicken MEF2D was not detected in lung and kidney of one week old chicken.

A summary of chicken MEF2D gene expression is given in Table 8.2.

Table 8.2 Summary and comparison of expression of MEF2D mRNA in one week and six weeks old chickens (NI: not investigated; +/-: expressed or unexpressed).

	Detection of mRNA	
Tissue	**1 week**	**6 weeks**
Leg Muscle	+	+
Breast Muscle	NI	+
Brain	+	+
Liver	+	+
Kidney	-	+
Ovary	+	+
Lung	-	NI
Heart	+	NI

Our results indicate that chicken MEF2D mRNA is ubiquitously expressed in both muscle and non-muscle tissues at different developmental stages. This mRNA expression pattern resembles that of the human and the *Xenopus* MEF2D gene. These results support the hypothesis that myocyte enhancer factor 2D acts as a regulator of myogenesis and neurogenesis.

Chapter 9

Discussion

9.1 Selection of Microsatellite Clones from Genomic DNA and cDNA

The copy numbers of microsatellites tend to increase with the haploid genome size in Yeast, *Drosophila Xenopus* and mammalian species, with exception of the chicken genome where the $[CA/TG]_n$ repeats are almost 10-fold less abundant than that found in mammals. Therefore for whole genome studies, it is necessary to construct microsatellite-enriched DNA libraries to provide sufficient microsatellite clones for screening.

In search of an efficient protocol for the rapid production of genomic libraries highly enriched for microsatellites, the "genetic marker selection technique" (Ostrander *et al.*, 1992) was used with some modifications. With chicken genomic DNA, the percentage of positive clones was 1.5% in the $[CA/TG]_n$-enriched library, and 0.5% in a $[CCT]_n$-enriched genomic library. This was about a 12-fold enrichment in the $[CA/TG]_n$-library and 8-fold enrichment in the $[CCT]_n$-library. However, further characterisation of the selected clones by DNA sequencing revealed that several clones were highly-represented in these libraries, reflecting a low complexity of the marker-selected libraries. This phenomenon was also observed by other groups (e.g. Cheng and Crittenden, 1993).

To overcome these limitations of the "marker-selected libraries", a "PCR-based DNA affinity hybridisation and capture procedure" was developed for the construction of microsatellite-enriched DNA libraries. In this procedure, a 5'-biotinylated microsatellite oligonucleotide is used to hybridise to heat-denatured $[CA]_n$-containing genomic DNA fragments. The heteroduplex molecules are physically captured and separated from the non-hybridised DNA molecules using streptavidin-coated magnetic beads. With this procedure, genomic libraries with up to 6% positive clones are obtained, which is about a 50-fold enrichment. This is a 4-fold increase over the marker-selected genomic library. A

comparison of the results obtained for the "marker-selected" and the "DNA affinity hybridisation-selected" libraries is given in Table 9.1.

Table 9.1 Comparison of the "genetic marker-selection" and the "DNA affinity hybridisation" protocols

Items Compared	**Genetic marker selection**	**DNA affinity hybridization**
Percentage of positive clones	Low	High
Complexity of the library	Low	High
Repeat motif length	Short	Long
Cloning steps involved	2	1
Amount of starting DNA	Large	Small
Manipulation	Complicated	Simple

DNA library generated with "DNA affinity hybridization" protocol has a higher percentage and higher complexity of positive clones than those generated by "marker selection" protocol. In addition, a very small amount of starting DNA is needed. Since all manipulations of this protocol are performed *in vitro* , the reaction conditions can be controlled more easily. Furthermore, only one cloning step is involved with the "DNA affinity hybridization" protocol. It is clear from this comparison that the "DNA affinity hybridisation" protocol provides a more efficient approach for the isolation of microsatellite from the chicken genome. Very recently, two further protocols for enriching microsatellites in genomic libraries have been reported (Armour *et al.*, 1994; Kandpal et al., 1994). These protocols are technically similar to the one developed in this study, in which DNA affinity hybridisation/capture is used to select for microsatellite-containing sequences. The difference lies in that, in Kandpal/Armour's methods, the microsatellite oligo nucleotide motif is fixed onto a nylon filter membrane.

The "DNA affinity hybridisation" protocol has provided an efficient way for enriching microsatellites in DNA libraries. However, when this protocol was used in construction of genomic or cDNA libraries, a number of shortcomings were found. In order to improve the efficiency of microsatellite enriching, the "DNA affinity hybridisation" protocol may subject to the following modifications .

A. Nearly 50% of the positive clones in our $[CA/TG]_n$-enriched genomic library contains [CA/TG]-rich sequences (TG-islands). Within these sequences, CA/TG repeats were scattered throughout the DNA insert with frequent interruptions by unique sequences. Construction of microsatellite-enriched DNA libraries using cDNA instead of genomic DNA showed that such [CA/TG]-rich sequences were less frequent (13% of the total positive clones). Therefore in the future, either different source of DNA should be used (e.g. cDNA), or DNA pre-selection techniques could be introduced, e.g. DNA re-association experiment or DNA subtractive hybridisation (Wang and Brown, 1991; Lisitsyn *et al.*, 1993) to eliminate such CA/TG-rich sequences.

B. In a proportion of the positive clones from the "DNA affinity hybridisation-selected" library, the repeating microsatellite motifs are found at the extreme ends of the cloned insert. Without further cloning experiments, it is difficult to design primers for PCR amplification from these clones. DNA sonication, the technique used for random-fractionation of high molecular weight genomic DNA in our experiments may be responsible for the generation of such clones. Sonication is expected to break DNA at random locations. In the future, it may be better to use frequent-cutting restriction enzymes such as *Alu I, RsaI* and *HaeIII* instead of sonication. The use of these enzymes will avoid cleavage of DNA within the [CA/TG] repeat motif, increasing the chances of generating suitable microsatellite clones (i.e. a repeat motif with sufficient flanking sequence for primer design).

C. In the current protocol, a single step of hybridisation and capture is used, which yielded 5% positives in the enriched library. In the future, if two or more hybridisation cycles are used, the proportion of non-specifically hybridised and captured DNA molecules will be significantly reduced. Such a library is expected to further improve the efficiency of large-scale isolation of microsatellites.

D. In the current protocol, oligo dT was used as the primer in the synthesis of first-strand cDNA from mRNA. However, using random hexamers enables all transcribed regions to be synthesized into cDNA. So microsatellites within all introns can be cloned through this protocol. In the future, random priming with random hexamers should be used in the generation of cDNA resource.

E. In the current protocol, two tandem repeat, a dinucleotide repeat (CA/TG) and a trinucleotide repeat (CAG/CTG) are used. In the chicken genome, there are other frequently distributed mono, di, tri, tetra and panta repeat motifs. Moran (1993) and Beckman *et al.* (1992), in separate studies, have compared the frequencies of microsatellite in pig, chicken, human and rat sequences found in the Genbank database. It was revealed that although no significant difference was found in the overall frequency of microsatellites in pig (13%) and chicken (10%), the distribution between repeat classes is quite different. As summarised in Table 9.3, the most pronounced difference is in the dinucleotide repeat category, especially the $[CA/TG]_n$-type dinucleotide repeats (Beckman and Weber, 1992; Moran, 1993). $[A]_n$ and $[CA/TG]_n$ are the two most frequently distributed microsatellites in the chicken genome. They are obviously good candidate repeats to start with in the construction of microsatellite-enriched genomic or cDNA libraries.

Table 9.3 Comparison of the observed frequency of microsatellite among different eucaryotic genomes (Beckman *et al.*, 1992; Moran, 1993)

Species	Overall (%)[a]	Dinucleotide Repeats (%)[b]	$[CA/TG]_n$ (%)[b]	Order of Abundance
Chicken	10	12	0.9	[A]>[AC]
Pig	13	30	3.9	[A]>[AC]
Human	-	28	19	[A]>[AC]>[AAAT]>[AG]>[AT]
Rat	-	53	36	[AC]>[AG]>[A]>[AAAT]>[AAGG]

a: number of microsatellites/number of gene entries
b: number of dinucleotide repeats/total number of microsatellites

Since very small quantities of starting DNA are required, a potential application of this approach is to construct chromosome-specific or region-specific microsatellite-enriched DNA library from micro dissected chromosomes. For several reasons, it is profitable to do so. First, with the accumulation of large numbers of microsatellites, continuing a random approach for microsatellites isolation would become increasingly unproductive. Second, in the later stages of genetic map construction, as the map reaches completion, there will exist areas poorly covered by polymorphic markers. Third, one of the targets of the human genome mapping project is to construct a high-resolution physical map in order to facilitate positional-cloning of genes mapped by linkage analysis. Such a goal is unlikely in livestock genomes. However, a local physical contig is more likely to be generated in order to clone mapped genes. For example, if a gene or QTL with important biological or economic value is mapped to a specific chromosome region, more markers may be isolated from this region to make a detailed map of this region prior to the final physical isolation of the gene.

A frequently asked question is that how many microsatellite markers are needed for chicken genome mapping ? The answer to this question varies depending on different goals of the chicken genome project. There are two distinct goals: the detection of trait-gene association and positional cloning of genes. Each goal demands different numbers of microsatellite markers.

For the detection of trait-gene association and measurement of gene effects, the strategy is a "whole genome analysis" (Todd *et al.*, 1994). In principle, chromosome-specific sets of microsatellite markers are used. These markers are evenly spaced throughout the genome with an average interval of about 10 cM. These markers are used to search for marker-trait associations in specific crosses. The total genetic size of the chicken genome has been estimated to be 2000 to 3000 cM (Levin *et al.*, 1994). Therefore to perform whole genome analysis, approximately 200 evenly-spaced microsatellites are required to produce an average space of 10 cM between markers. Microsatellites are unlikely to be evenly dispersed throughout the chicken genome. For example, in isolating and mapping human microsatellites, a proportion of the microsatellites were clustered in one or more regions (Weissenbach *et al.*, 1992). Therefore in practice, a minimum of 1000 random microsatellites need to be mapped to allow one to identify 200 evenly spaced markers to be used as index markers to facilitate such a whole genome analysis (Todd *et al.*, 1994).

For map-based cloning of candidate genes for specific traits, a high-resolution genetic linkage map (1 cM) and a high density physical map using YACs (1 marker/0.5-1.0 Mb) are needed. Such a goal is unrealistic in chicken. An achievable target will be the construction of high resolution genetic linkage maps and a local physical map once a gene or QTL is mapped to a specific region. The number of markers required will depend

on the genetic distance between the two closest flanking markers, between which the candidate gene is mapped.

Currently, eight collaborating groups (Europe and USA) are isolating microsatellite markers for the chicken genome project. The development of the microsatellite-enriching protocols for constructing microsatellite-enriched DNA libraries in this thesis will facilitate the rapid isolation of large numbers of microsatellite markers for the chicken genome mapping project.

9.2 Trait Gene Identification

One of the long-term goals of the chicken genome mapping project is to identify individual genes which control a biological or economic trait e.g. production, reproduction and health traits. Conventionally, two approaches have been used: comparative mapping and positional cloning. The comparative mapping approach can be used in chicken, but, the positional cloning approach seems unrealistic because of the lack of high resolution physical maps in chicken. Instead, a whole genome analysis approach is developed in the identification of trait genes in chicken.

9.2.1. Candidate gene identification via comparative mapping

Since the divergence of all the species from their common ancestor, the arrangements and the number of chromosomes have changed due to the processes such as chromosome translocations, insertion/deletions and fusions. However, it was often found chromosome regions identical among different species in terms of gene loci. This is more common for

species that share a recent ancestor. Table 9.4 shows examples of conserved syntenies found among vertebrates (Burt, unpublished data). Since gene mapping in human and mouse are more detailed and systematic than in other species, a comparison of the human/mouse gene map with that of the chicken can lead to the proposal of candidate genes within an homologous chromosome region.

Table 9.4 Homologous linkage groups in vertebrates

Locus	Human	Chick	Cow	Pig	Sheep	Mouse	Rat
HBB	11	1-2	15	-	-	7	1
PGR	11	1	-	-	-	9	-
GAPD	12	1	5	-	3	6	-
LDHB	12	1	5	5	3	6	4
IGF1	12	1	5	5	3	10	7
LYZ	12	1	5	5	3	10	-
H1F4	12	1	-	-	-	13	-
ACTB	7	2	-	-	-	5	-
EGFR	7	2	-	-	-	11	14
CA2	8	2	14	-	9	3	-
CALB1	8	2	-	-	-	4	-
MYC	8	2	14	-	9	15	7
YES1	18	2	24	-	-	-	1
BCL2	18	2	-	-	-	1	-
							-

* Human locus symbols used; Human data taken from GDB (Sept 1994); Mouse data taken from MGD (Release 1.0, Sept 1994); Chicken data taken from CHICKBASE (July 1994).

By screening the microsatellite-enriched genomic or cDNA libraries generated in this study, an increasing number of microsatellite markers will be identified. This steady accumulation of large numbers of both genomic and cDNA microsatellites in the chicken allows one to use these microsatellites in the mapping of the other avian genomes. Recent studies have tested the validity of using heterologous primers in mapping related species (Moore *et al.*, 1991; Stallings *et al.*, 1991). This should be possible with poultry. A very recent study shows that at least 50% of the chicken microsatellite markers can be used for

genome mapping in the turkey (Levin *et al.*, personal communication). This will not only save the overall cost and time for developing microsatellite maps in these species, but will also allow the transfer of mapping information from the map-rich species (e.g. chicken) to the map-poor species (e.g. turkey).

More importantly, the cDNA microsatellites identified in this study represent expressed genes, so that they can be classified as type I loci (conserved loci). The mapping of these cDNA microsatellites will allow one to compare genetic maps between distantly related species (e.g. human and chicken). Consequently, map information will be transferred from the human genetic map to the chicken. Such a comparison of the genetic maps between chicken and mammalian species may also lead to the identification of candidate genes for a specific biological or economic trait in poultry.

9.2.2. Candidate gene identification via whole genome analysis

The goal of a whole genome analysis is the identification of chromosome regions (loci) that control a specific biological phenotype or production trait. In the long term, the chromosome region will have to be cloned to facilitate further characterisation of the gene. There are two possible routes to follow: First, to construct a region-specific high resolution genetic linkage map from the analysis of further markers and genetic crosses. This map can be used to further define the location of the trait gene. The next step is to use flanking markers to isolate YAC clones. A physical contig map made of overlapping YAC and cosmid clones may then allow the identification of the genes within the region. This is a very specified and targeted approach. Second, with the assigning of large numbers of microsatellite-containing cDNA sequences to the mapped region, some of these expressed sequences may be candidate genes for that mapped trait. Further

investigations on other characteristics of these mapped expressed sequences, e.g. expression patterns or possible functions, are required to test out this hypothesis. Furthermore, the chance of finding a candidate gene may be increased by choosing appropriate RNA sources. For example, to identify candidate genes responsible for muscle development, a good RNA source to choose to make cDNA library is the muscle satellite cells.

Microsatellites are highly polymorphic, and evenly distributed throughout the genome. They are suitable for developing high-resolution genetic maps. Therefore these markers are used in the "whole genome scanning" strategy to map chromosome regions that control biological or economic traits. These markers can also be used in constructing high density physical maps by simply ordering large-insert clones such as YACs or cosmid clones. Once a high-resolution genetic map and a high-density physical map have been developed, positional cloning can be conducted to 'walk' to the target chromosome region.

Through comparative mapping and "whole genome scanning", a number of genes or chromosome regions are identified. With the availability of these candidate genes for a specific trait, further studies may focus on the comparison of their expression patterns, gene structures and evolutionary conservation, sequence homology to known sequences in the data bases, possible protein domains and the function of the protein, mutations and function. These will test the "candidate gene" hypothesis. For example in tomato, a candidate cDNA clone was mapped to the genomic region that confers resistance to a pathogen, *pseudomonas syringae pv.tomato.* When susceptible tomato plants were transformed with the cDNA, they were resistant to the pathogen, strongly implicating that this cDNA clone is the disease resistant gene. Analysis of the amino acid sequence

revealed similarity to serine-threonine protein kinases (Martin and Brommonschenkel *et al.*, 1993).

9.2.3. Other strategies for identification of candidate genes

Trapping of candidate genes that control monogenic or polygenic traits can also be facilitated by randomly isolating and mapping large numbers of expressed sequences (cDNA).

A. Expressed sequences are conserved loci. They can be used in comparative mapping between distantly-related species (e.g. human/mouse and chicken) together with cloned genes.

B. Expressed sequences can also be used directly as additional markers to map onto the chromosome region identified by the "whole genome scanning" process using specific line crosses. If the established linkage data for a particular trait or disease condition implicates a specific chromosomal region, the availability of several cDNAs known to map to that region would provide a number of candidate genes for that trait or disease. For example, the Wilson disease (WD) in human was mapped to a single marker interval at 13q14.3 (Petrukhin *et al.*, 1993). Recently, a partial cDNA clone (pWD) was also mapped to this region (Tanzi *et al.*, 1993). Detailed analysis of the predicted functional properties of the pWD gene and the identification of four independent disease-specific mutations provide convincing evidence that pWD is the Wilson disease gene.

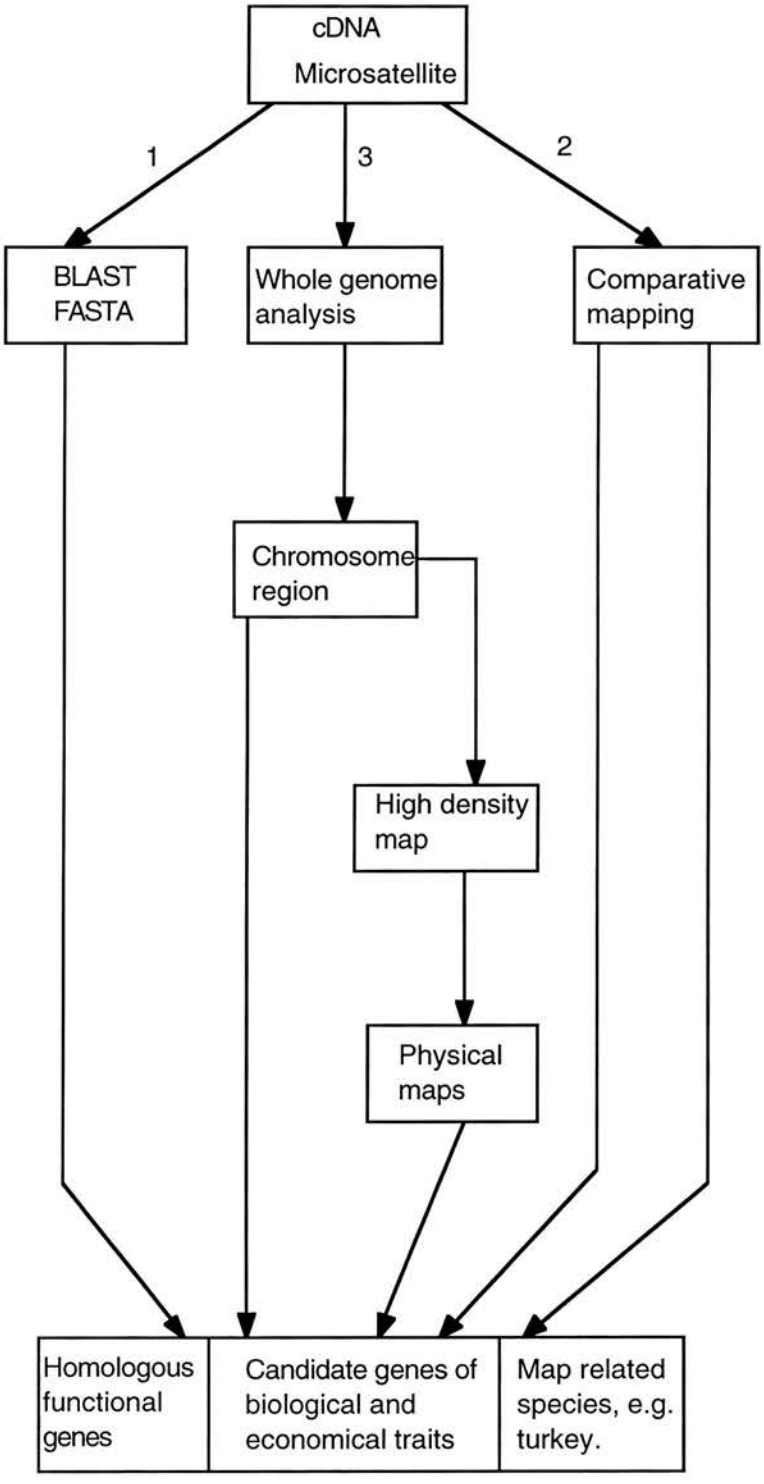
The current strategies used to map expressed sequences are: (1) RFLP analysis of genomic DNA using anonymous expressed sequences as probes (Bumstead and Palyga, 1992); (2) generation and analysis of expressed sequence tags (ESTs) (Wilcox *et al.*,

1991). Both methods have been successful, in particular, the "expressed sequence tag" approach, since the PCR technique has been integrated in the method. However, the shortcomings with these methods are slow and high cost.

To identify candidate genes for physiological and production traits, large numbers of expressed sequences have to be mapped. This demands a fast method for mapping expressed sequences. Isolating microsatellite-containing cDNA and use of the microsatellites to map these expressed sequences is a potential approach.

We have constructed a $[CA/TG]_n$-enriched cDNA library using the DNA affinity hybridisation and capture method developed for genomic DNA sequences. Screening of the library has shown that $[CA/TG]_n$ microsatellites exist in a wide variety of expressed sequences. Most of these $[CA/TG]_n$ microsatellites are located in either the 5' or 3' untranslated regions. However, we have evidence from the alignment of hFMR1 and contig CA003, PLT and contig CA023 DNA sequences, to argue that some of these microsatellites may come from introns, which may be due to the presence of some pre-processed RNA in the mRNA preparations.

Whatever their locations within genes, these cDNA microsatellites, in contrast to the genomic microsatellites which represent anonymous sequences of unknown function, represent expressed genes. Therefore, these cDNA microsatellites have various potential usage, as illustrated in the following figure.



Applications of cDNA microsatellites

9.3 Characterization of Homologous Genes

Nucleotide sequencing of cDNA clones has generated a large number of random sequences. This nucleotide sequence information of the microsatellite-containing cDNAs can be used to perform database similarity searches using BLASTX, BLASTN, FASTA programs to detect homologous sequences. Using this strategy, we have defined 4 chicken homologues of known human genes and a further 4 sequences related to other genes in other species, e.g. human, yeast and *E. coli*. This strategy has made it possible to link yeast and mammalian genetics to the chicken genome via identification and mapping of the chicken homologues.

9.3.1. The Chicken Homologue of Human Fragile X Mental Retardation Syndrome Gene (FMR1)

A partial cDNA fragment (CA003) was cloned by screening a $[CA/TG]_n$ microsatellite-enriched liver cDNA library. A non-redundant database similarity search using the BLASTX program showed that this partial cDNA fragment is homologous to the human fragile X mental retardation syndrome gene, hFMR1. DNA sequence comparison between hFMR1 and CA003 revealed a sequence similarity of 85% over a 176 bp DNA region.

The cDNA fragment, CAG003, consists of 537 bp. In examination of the sequence alignment of CAG003 and hFMR, and the sequences at the last intron-exon border region of hFMR1 gene, it was found that CAG003 can be divided into two parts. The region between nucleotide 362 and 537 corresponds to exon XVII of hFMR1, whereas the first

362 nucleotides of the fragment is located in the last intron. This is an indication that CAG003 was cloned from an unprocessed mRNA.

A $[CA/TG]_{14}$ microsatellite is found 208 bp upstream of the determined intron-exon border in the cloned partial chicken FMR1 gene. The identification of such a microsatellite will provide a polymorphic marker to rapidly map this FMR1 gene onto the current chicken genetic groups.

Southern hybridisation using human FMR1 probe to screen the DNAs from a number of different species including Chimpanzee, Gorilla, Rhesus, Chicken, *Drosophila*, Nematode and Yeast has revealed extensive cross-species conservation of this gene except in *Drosophila* (Verkerk *et al.*, 1991). Such a high cross-species conservation suggest that FMR1 gene plays an important role *in vivo*. Indeed, studies have shown that the murine FMR1 gene was transcribed in a ubiquitous manner with an expression pattern similar to glyceraldehyde phosphate dehydrogenase (GAPDH) with an enhanced expression during germ cell proliferation (Bächner *et al.*, 1993). In human, FMR1 is found transcribed widely during development (Hanzlik *et al.*, 1993).

A further question to be addressed is what actual role does the FMR1 protein play *in vivo* ? Two recent studies have reported the identification of two domains within the FMR1 protein by amino acid sequence comparison, the RGG and KH domains which are usually found in the family of RNA binding proteins (Siomi *et al.*, 1993; Ashley *et al.*, 1993; Burd and Dreyfuss, 1994). In these studies, the FMR1 protein was found bound to a wide range of RNA molecules including its own message. These results suggest that the protein product of the FMR1 gene may serve to transport and localise RNAs.

The cDNA fragment, AC003, obtained in this study could be used as a probe to screen a chicken genomic or cDNA library to isolate the complete genomic or cDNA clone for chicken FMR1. Then nucleotide sequence and amino acid sequence comparisons can be made between chicken and other species. This will lead to the identification of conserved motifs within the gene and protein, and confirm our conclusion that CA003 is the chicken FMR1 homologue.

The $[CA/TG]_{14}$ microsatellite identified within the last intron of the chicken FMR1 gene can be used as a genetic marker. Based on the mapping resource populations (East Lancing and Compton), the chicken FMR1 gene should be rapidly assigned to the current genetic linkage map.

9.3.2. The Chicken Homologue of Mouse NBL4

Several proteins contain the domain homologous to the N-terminal half of the "Band 4.1" protein and form a superfamily. The members of this family are thought to play crucial roles in the regulation of cytoskeleton-plasma membrane interaction (Takeuchi *et al.*, 1994). A non-redundant database search using the BLASTX program has detected sequence homology between MMNBL4 and chicken cDNA sequence, CA011. MMNBL4 and CA011 have a sequence similarity of 72% over a 377 bp DNA.

Examination of the sequence alignment of CA011 and MMNBL4, a number of base substitutions and deletion/insertions are found. In checking the original sequence, it was confirmed that these are not due to sequencing errors. Such deletion/insertions will cause frame shifts in the coding region. Moreover, a stop codon is found at the end of the open reading frame of MMNBL4, while stop codons are frequently located within the 3'

untranslated region in all three reading frames. Together, these observations suggest that the homologous region between CA011 and MMNBL4 is highly unlikely to be translated. This conclusion is in agreement with the result of a recent study (Takeuchi *et al.*, 1994). The deduced amino acid sequence of NBL4 revealed a myristorylation site, as well as phosphorylation sites for A-kinases in its N-terminal half, suggesting its involvement in the phosphorylation-dependent regulation of cellular events.

In this study it is found that as high as 72% DNA sequence homology is seen between the chicken cDNA clone, CA011, and the 3' untranslated region of MMNBL4. This high level of homology in the 3' untranslated region between mouse and chicken strongly suggests that this region is very important. Possibly it may be involved in the maintenance of NBL4 mRNA stability and/or translational control of the expression of the NBL4 gene. For example, Aly *et al.* (1994) observed that the 3' untranslated region of a heat-shock protein gene in *Leishmania* , hsp83, is involved in the temperature-dependent degradation of its mRNA. It is possible that this region may form some secondary structures, since DNA sequence homology between chicken and mouse homologues is seen in a long region of the 3' untranslated region. Further studies are needed to test this hypothesis.

For a better understanding of the functions of NBL4 and other member of the "Band 4.1" superfamily, it would be necessary to clone the full length cDNA of the chicken NBL4. The cDNA clone identified in this experiment, CA011, can be used as probe. This will allow a full sequence comparison between the mouse NBL4 and its chicken homologue which may lead to the highlight of some conserved motifs.

A sequence comparison between the mouse and chicken NBL4 whole cDNA will be able to define limits of the sequence homology. For example, whether the similar sequence homology exists in the coding region or only the 3' untranslated region? It is interesting to see whether this conserved region is a 3'UTR regulatory sequence present in many other genes?

9.3.3. The chicken homologue of human PLT

Non-redundant database search using the BLASTX program showed that cNDA clone CA023 is homologous to a human sequence, hPLT. Sequence comparison revealed a sequence similarity of 69% over a 130 bp DNA. hPLT is a functionally-unknown human gene containing the long terminal repeat (LTR) of the RTVL-H family of human endogenous retroviral-like elements (Goodchild *et al.*, 1992).

Examination of the sequence alignment between hPLT and CA023 detects sequence homology between nucleotide 211 and 340 in CA023, and the corresponding region in the human PLT. However, there is no significant sequence homology with CA023 from nucleotide 1 to 210. Further examination of the conserved boundary reveals a possible consensus splice acceptor site with the CA023. This again suggests the cloning of a pre-processed mRNA.

Within the intron region identified, a $[CA/TG]_{14}$ microsatellite is located 209 bp upstream of the intron-exon border. This microsatellite can be used as polymorphic marker to map the chicken PLT gene into genetic linkage map

9.4 [CAG/CTG]$_n$ Trinucleotide Repeat and Transcriptional Factors

The original purpose of constructing microsatellite-enriched DNA libraries is to provide better resources for screening, in order to improve the efficiency of isolation of microsatellite markers. However, recently, a number of interesting studies showed that [CAG/CTG]$_n$ trinucleotide repeats were frequently associated with transcriptional factors or other developmentally important genes.

Homopolymeric stretches of amino acids are found in an increasing number of proteins (Chamberlain *et al.*, 1994). Gerber *et al.* (1994) searched the Swiss-Prot protein database for polymeric stretches of at least 20 glutamines (Q) and 10 prolines (P). He found 33 out of the 40 top-scoring proteins having polymeric glutamine stretches (82%) were transcription factors. Similarly, out of the 44 top scoring proteins bearing a poly-proline stretch, 34 (78%) were thought to be transcriptional factors. A selection of the transcriptional factors found was shown in Table 9.5.

The existence of poly-glutamine tracts in a proportion of transcriptional factors indicates that they may be important to the function of these transcriptional factors. Based on a protein fusion experiment, Seipel and colleagues proposed that the glutamine-rich domain acts as "Proximal" activation domain (Seipel *et al.*, 1992). Poly-glutamine or glutamine-rich stretch is usually used to contact components of the general transcriptional apparatus. A possible explanation is that glutamine stretch or glutamine-rich domain is rarely charged, and is hydrophobic. Such structure properties may favour the attachment of transcription factor to cis-acting elements and interaction with other proteins (Duboule *et al.*, 1987). Another possibility is that the amide moieties of the glutamine side chains are

involved in hydrogen bonding to RNA polymerase or some other components of the general transcriptional machinery (Courey and Tjian, 1988).

Table 9.5 Transcriptional factors with polymeric glutamines found in the Swiss-Prot protein database (Yu *et al.*, 1992; Breitbart *et al.*, 1993; Gerber *et al.*, 1994)

Transcription Factors	**Length of Q-stretch (residues)**	**Q within Stretch (%)**	**Organism**
Myocyte enhancer factor 2D	40	62	Human
Embryonic polarity dorsal protein	34	86	*Drosophila*
Androgen receptor	21	100	Human
Transcription protein GAL11	23	100	Yeast
Nuclear protein SNF5	37	100	Yeast
TATA binding protein (TBP)	38	100	Human
Transcriptional activator DAL81	22	95	Yeast
Grainy head	23	96	*Drosophila*
Hunchback protein	24	96	*Drosophila*
Huntingtin	11-34	100	Human
Glucocorticoid receptor (GR)	22	96	Rat
Suppresser 2 of Zeste protein	26	96	*Drosophila*
Neurogenic protein mastermind	20	95	*Drosophila*
Antennapedia homeotic protein	29	76	*Drosophila*
Neurogenic locus Notch protein	31	97	*Drosophila*
Octamer transcription factor N-OCT3	25	96	Human
Homeotic sex combs reduced (SCR)	20	80	*Drosophila*

Poly-glutamine tracts are encoded by repetition of the codon [CAX] (X=A, G), forming a trinucleotide repeat (microsatellite). Theoretically, if a $[CAG]_n$ triplet is used to construct

a [CAG]-enriched cDNA library, a proportion of the positive clones should be chicken transcriptional factors. Using this strategy, we have obtained 7 $[CAX]_n$-containing cDNA sequences by screening a $[CAG]_n$-enriched chicken liver cDNA library. CAG001 and CAG007 contain a single run of 43 and 37 [CAX] repeats, respectively. CAG009, CAG015 and CAG020 contain complex triplet repeats, in which the $[CAX]_n$ repeat is interrupted by other trinucleotides. In CAG018 and CAG019, the [CAX] trinucleotides are scattered along the whole cDNA sequence. In an examination of the nucleotide composition of these triplet cDNA sequences, we found that, the first two nucleotides within each repeat unit remains constant, whereas the third nucleotide could either be A or G. These results indicate that this sequences are under natural selection.

CAG001 is a 567 bp chicken cDNA fragment. This partial cDNA contains a contiguous stretch of 43 [CAX] trinucleotide repeats (X=A or G). A non-redundant database similarity search using the BLASTX program indicates that this partial cDNA is homologous to the human myocyte enhancer factor 2D (MEF2D).

Myocyte enhancer factor 2 (MEF2) was originally identified from muscle nuclear extracts. It binds the A/T-rich sequence motif in a number of muscle-specific enhancers and promoters to modify the expression of these genes.(Martin *et al.*, 1993; Yu *et al.*, 1992; Gossett *et al.*, 1989). Various MEF2 factors and their corresponding cDNAs have been cloned in both human and other species (Martin *et al.*, 1993; Chambers *et al.*, 1992; Yu *et al.*, 1992; Pollock *et al.*, 1991; McDermott *et al.*, 1993). Multiple isoforms of human MEF2 proteins are the products of at least four genes, hMEF2A, hMEF2B, hMEF2C and hMEF2D. Characteristically, all the factors have two DNA binding domains at the N-terminus of their proteins, the MADS domain and the MEF2-specific domain. These factors therefore comprise a gene family, termed the MEF2 family. Since

all these factor share a highly conserved MADS domain with a number of homeotic genes in plants and yeast, they also belong to the MADS super family.

9.4.1 The poly-(Q/P) tract is a potential transcriptional activation domain

Poly-glutamine/proline tracts are found in a large proportion of the transcriptional factors (see introduction). Experimental evidence suggests that these poly-(Q/P) tracts might act as transcriptional activation domains (Gerber *et al.*, 1994).

We have examined the poly-[CAX] tract in the cloned partial chicken MEF2D cDNA, and found that the first two nucleotides within each repeat units remain constant, whereas the third nucleotide is either A or G. However this change will not result in an amino acid difference, since both CAA and CAG code for a glutamine residue.

We have also compared the poly-(Q) tract and its flanking regions in the chicken MEF2D partial peptide sequence with that of the human MEF2D peptide. As shown in Figure 6.7, within the poly-(Q) tract we find a deletion/insertion of 2 glutamine residues and a point mutation (Q to S), whilst the rest of the poly-(Q) tract remains highly conserved. One striking observation is that in chicken, the poly-(Q) tract is uninterrupted, whereas in human it is a mixture of Q and P. This suggest that glutamine and proline residues are inter-changeable, assuming that MEF2D proteins have identical functions in chicken and human. High conservation is also seen at the 5' flanking region of the poly-(Q) tract. Given that chicken and human have diverged from a common ancestor about 300 million years ago, our results suggest that the poly-(Q/P) tract and its 5' flanking region may be important in the function of these genes: possibly as transcriptional activation/inhibition domain as suggested in previous studies (Gerber *et al.*, 1994; Chamberlain *et al.*, 1994).

9.4.2. The $[CAX]_n$ triplet in the MEF2D gene is a mutation "hot spot"

A large number of studies have shown that microsatellites are highly-unstable sequence elements. In human, seven neurodisorders have been correlated with the dynamic expansion of triplet repeats within the corresponding genes.

A contiguous run of 43 glutamine residues is found near the C-terminus of the chicken MEF2D partial peptide sequence. Similar elements are also found in human and *Xenopus* MEF2D peptide sequences. As shown in Figure 6.7, while the flanking amino acid sequences remain highly conserved, there are differences within the poly-glutamine element. A deletion/insertion of 2 glutamine residues had occurred since the divergence of the chicken and the human MEF2D genes. More changes are seen with *Xenopus* : a deletion/insertion of 29 glutamine or proline residues, together with several point mutations. These results strongly support the hypothesis that trinucleotide repeats in genes are undergoing dynamic mutations which lead to deletions or insertions.

9.4.3. Ubiquitous expression of MEF2D mRNA in Chicken Tissues

Previous studies have shown that the mRNA of both the human MEF2D gene and its *Xenopus* homologue, SL-1, was ubiquitously expressed in a wide variety of tissues (Yu *et al.*, 1992; Breibart *et al.*, 1993). We have investigated the pattern of expression of the chicken MEF2D gene using RT-PCR. A comparison of the expression patterns of the human, *Xenopus* and chicken MEF2D mRNA is summarised in Table 9.7.

Our investigations show that the chicken MEF2D mRNA is also ubiquitously expressed in both muscle (e.g. leg muscle, breast muscle and heart) and non-muscle (liver, ovary, brain and kidney) tissues. This expression pattern is consistent with that found in human

and *Xenopus* . Although post-transcriptional regulation might be important in the generation of tissue-specific activity for the MEF2D gene, the ubiquitous pattern of expression of human, chicken and *Xenopus* mRNA suggests that it may play a "house-keeping" role *in vivo*.

Table 9.7 Comparison of the expression patterns of human, *Xenopus* and chicken MEF2D mRNA.

	Chicken	**Human**	***Xenopus***
Heart	+	+	+
Skeletal Muscle	+	+	+
Brain	+	+	+
Liver	+	-	+
Kidney	+/-	-	+
Ovary	+	ND	+
Lung	-	+	ND
Placenta	ND	+	ND
Spleen	ND	ND	+
Stomach	ND	ND	+

ND, Not done; +, expressed; -, unexpressed.

Similar to MEF2D, the mRNA of both human MEF2A and its *Xenopus* homologue, SL-2, also display a ubiquitous expression pattern (Yu *et al.*, 1992). However, in contrast to the MEF2A and MEF2D genes, MEF2B and MEF2C showed a restricted expression pattern. Their mRNA was largely restricted to skeletal muscle and brain (McDermott *et al.*, 1993). This difference is also reflected in their amino acid sequence. For example, a highly conserved poly(Q/P) tract is found in both MEF2A and MEF2D gene, but not in MEF2B and MEF2C gene. Previous studies showed that these poly-(Q/P) elements might

function as "Proximal" activation domain which might allow efficient channelling of an enhancer effect to a responsive promoter (Seipel *et al.*, 1992).

9.4.4. Further studies

Based on the initial work on the MEF2 gene family reported in the thesis, the following areas are interesting for further investigations.

A. Identifying and cloning other members of the MEF2 family. The cDNA fragment obtained in this experiment, CAG001, can be used as probe to clone its full length cDNA. Further sequence comparison will verify the MEF2D homology in the rest of the protein. A full length sequence will allow identification and comparison of the promoter sequences of the gene. cDNA probes can be designed across the conserved MADS-domain or the MEF2-domain region and used to probe cDNA libraries to pull out other members of the MEF2D family. A genomic southern blot can be done using the cloned MEF2D gene cDNA as probe to determine how many genes within the MEF2 gene family exist in the chicken genome. With the availability of mapping crosses, the isolated genes could also be mapped onto individual chromosomes or in linkage groups.

B. Detailed investigation of expression patterns. In this study, gene expression of the chicken MEF2D gene was only looked at the mRNA level. Since post-transcriptional regulation has been demonstrated to be important in the generation of tissue-specific MEF2 activity, it is necessary to investigate the protein of these genes by immunohistochemical assays prior to defining functions for these genes.

C. Gene regulation studies. With the cloning of the chicken MEF2 genes, *in vitro* gene expression studies using reporter genes could then be performed to study expression regulation for these genes.

When the sequences of clones CAG007, CAG009, CAG015, CAG018, CAG019 and CAG020 were used in database searches, no homologous genes were identified. To predict whether these undefined $[CAX]_n$-containing cDNAs have any coding potential, we analysed these sequences using GRAIL1 (Gene Recognition and Analysis Internet Link) (Uberbacher and Mural, 1991). All of the sequences analysed were predicted to contain long open reading frames (ORF). Potential exons have been recognised in these sequences with a quality score of 'good' or 'excellent'. These results indicate that the $[CAX]_n$-containing cDNA sequences obtained in our study are probably parts of the coding regions of their corresponding genes. Since no sequence homology has been detected in the current databases, it is highly likely that these clones may represent novel $[CAX]_n$ triplet-containing genes. Previous studies showed that "opa"-containing genes are differentially regulated during development and probably encode transcriptional factors (Wharton *et al.*, 1985; Gerber *et al.*, 1994). Based on these results, we propose the [CAX]-containing cDNA clones identified in our study may belong to genes that encode transcriptional factors or other developmentally important genes. Further investigations can be carried out in the following areas:

A. Cloning and characterisation of the whole cDNA. The "opa" element-containing clones obtained in our experiments are all potentially located within coding regions of genes. To further understand the function of these genes, cloning of full length cDNAs and determination of their structure are essential.

B. Identification and cloning of novel triplet-containing genes. In order to isolate more opa-containing genes, different sources of chicken mRNA could be screened, e.g. brain, muscle, and ovary. Alternatively, mRNA of a tissue or whole embryo at different developmental stages could be screened. These will lead to the identification of a large number of potential developmentally-regulated genes or tissue-specific transcription factors.

C. Investigation of expression. Northern blotting or RT-PCR could be used to test if the genes are actually transcribed. If they are transcribed, it will be interesting to see where and when and what kind of role they play *in vivo*.

References

Aly, R., Miriam Argaman, Segula Halman and Michal Shapira (1994) *Nucleic Acids Research* **22**, 2922-2929.

Altschul, S. F., W. Gish, W. Miller, E.W. Myers and D.J. Lipman (1990) *Journal of Molecular Biology* **215**, 403-410.

Andersson, L., ChrisS. Haley, HansEllegren *et al.* (1994) *Science* **263,** 1771-1774.

Armour, J.A., Rita Neumann, Stephanie Gobert and Alec J.Jeffreys (1994) *Human Molecular Genetics* **3**, 599-605.

Ashley, C.T., Keith D.Wilkinson, Daniel Reines and Stephen T.Warren (1993) *Science* **262**, 563-566.

Avner, A., L. Amar, L. Dandolo and J. L. Guenet (1988) *TIG* **4**, 18-23.

Bächner, D., Antonella Manca, Peter Steinbach, Doris Wöhrle, Walter Just, Walther Vogel Horst Hameister and Annemarie Poustka (1993) *Human Molecular Genetics* **2**, 2043-2050.

Banerji, J., Laura Olson and Walter Schaffner (1983) *Cell* **33**, 729-740.

Banfi, S., Antonio Servadio, Ming-yi Chung *et al.* (1994) *Nature Genetics* **7**, 513-519.

Beckmann, J.S. and James Weber (1992) *Genomics* **12**, 627-631.

Beier, D.R. (1993) *Mammalian Genome* **4,** 627-631.

Bernardo Nadal-Ginard (1993) *Development* **118**, 1095-1106.

Bernúes, J., R. Beltrán and F. Azorín (1991) *Gene* **108**, 269-274.

Bio-Rad Instruction Manual Catalog Number 170-3581 *Muta-Gene Phagemid In Vitro Mutagenesis Version 2*

Bishop, M., Steven M. Kappes, John W. Keele, Roger T. Stone, Sara L. F. Sunden, Gregory A. Hawkins, Sabina Solinas Toldo, Ruedi Fries, Michael D. Grosz, Jakyoung Yoo and Craig W. Beatie (1994) *Genetics* **136,** 619-639.

Blackburn, E.H. (1991) *Nature* **350**, 569-573.

Blaho, J.A. and Robert D. Wells (1989) *Progress in Nucleic Acid Research* **37**, 107-126.

Breibart, R. E., Chang-seng Liang, Leslie B. Smoot, Dan A. Laheru, Vijak Mahdavi and Bernardo Nadal-Ginard (1993) *Development* **118**, 1095-1106.

Britten, R. J. and D. E. Kohne (1968) *Science* **161**, 529-540.

Bronner, C.E., Sean M. Baker, Paul T. Morrison *et al.* (1994) *Nature* **368**, 258-

Bullock, P., Josh Miller and Michael Botchan (1986) *Molecular and Cellular Biology* **6**, 3948-3953.

Bumstead, N. and Jan Palyga (1992) *Genomics* **13**, 690-697.

Burd, C.G. and Gideon Dreyfuss (1994) *Science* **265**, 615-621.

Burt, D.W., Jaquie M. Boswell, Ian R. Paton and Simon C. Butterwith (1992) *Biochemical and Biophysical Research Communications* **187**, 1298-1305.

Burt, D.W. and I. R. Paton (1992) *DNA and Cell Biology* **11**, 497-510.

Burt, D.W. and Sonia B. Jakowlew (1992) *Molecular Endocrinology* **6**, 989-992.

Burt, D.W., I. R. Paton and B. R. Dey (1991) *Journal of Molecular Endocrinology* **7**, 175-183.

Chamberlain, N.L., Erika D. Driver and Roger L. Miesfeld (1994) *Nucleic Acids Research* **22**, 3181-3186.

Chambers, A.E., S. Kotecha, N. Towers and T. J. Mohun (1992) *The EMBL Journal* **11**, 4981-4991.

Cheng, H.H., Lyman B. Crittenden (1993) *Poultry Science* 539-546.

Conley, E.C. and Stephen C. West (1989) *Cell* **56**, 987-995.

Courey, A.J. and Robert Tjian (1988) *Cell* **55**, 887-898.

Crittenden, L.B., Leonard Provencher *et al.* (1993) *Poultry Science* **72,** 334-348.

Crittenden, L.B. (1993) *Arch. Geflugelk* **57**, 59-68.

Crittendem, L.B., Leonard Provencher, Lisa Santangelo, Ilan Levein, Hans Abplanalp, Ruth W. Briles, Elwood Briles and Jerry B. Dodgson (1992) *Poutry Science* **72,** 334-348.

Crooijmans, R.P., A. J. A. Van Kampen, J. J. Van der poel, M. A. M. Groenen (1993) *Animal Genetics* **24**, 441-443.

Currier, R.D., G. Glover, J. F. Jackson and A.C. Tipton (1972) *Neurology* **22**, 1040-1043

Derynck, R., Lucy Rhee, Ellson Y.Chen and Ann Van Tilburg (1987a) *Nucleic Acids Research* **15**, 3188-3189.

Derynck, R. and Lucy Rhee (1987b) *Nucleic Acids Research* **15**, 3187.

Derynck, R., Patricia B. Lindquist, Angela Lee *et al.* (1988) *The EMBL Journal* **7,** 3737-3743.

Dey, B. (1992) Transforming Growth Facot-ß3 Gene-Structure and Evolution, *PhD Thesis*, University of Edinburgh.

Dietrich, W., Hillary Katz, Stephen E. Lincoln, Hee-Sup Shin, Jeffrey Friedman, Nicholas C. Drcopoli and Eric S. Lander (1992) *Genetics* **131**, 423-447.

Duboule, D., Marc Haenlin, Brigitte Galliot and Eliane Mohier (1987) *Molecular and Cellular Biology* **7**, 2003-2006.

Eichler, E.E., Stephen Richards, Richard A. Gibbs and David L.Nelson (1993) *Human Molecular Genetics* **2**, 1147-1153.

Estoup, A., Pablo Presa, Francine Krieg, Daniel Vailman and Rene Guyomard (1993) *Heredity* **71**, 488-496

Fields, C., Mark D. Adams, Owen White and J. Craig Venter (1994) *NatureGenetics* 7, 345-346.

Fishel, R.A., Kristina Detmer and Alexander Rich (1988) *Proc. Natl. Acad. Sci. USA* **85**, 36-40.

Fishel, R., Mary Kay Lescoe, M.R.S. Rao *et al.* (1993) *Cell* **75**, 1027-1038.

Flanagan, J.G. Marie-Paule Lefranc and Terence H. Rabbitts (1984) *Cell* **36**, 682-688.

Fornage, M., Lawrence Chan, Gerard Siest and Eric Boerwinkle (1992) *Genomics* **12**, 63-68.

Fu, Y.H., A. Pizzuti, R. G. Fenwick, Jr. J. King, S. Rajnarayan, *et al.* (1992) *Science* **255**, 1256-1258.

Fu, Y.H., Derek P.A. Kuhl, Antonio Pizzuti *et al.* (1991) *Cell* **67**, 1047-1058.

Gaillard, C. and Francois Strauss (1994) *Science* **264**, 433-436.

Gerber, H.P., Katja Seipel, Oleg Georgiev, Manuela Höfferer, Martin Hug, Sandro Rusconi and Walter Schaffner (1994) *Science* **263,** 808-811.

Gillies, S.D., Sherie L. Morrison, Vermon T. Oi and Susumu Tonegawa (1983) *Cell* **33**, 717-728.

Glick, A.B. *et al.* (1991) *Development* **111**, 1081-1086.

Goodchild, N.L., David A. Wilkinson and Dixie L. Mager (1992) *Gene* **121**, 287-294.

Gossett, L.A., David J. Kelvin, Edward A. Sternberg and Eric N. Olson (1989) *Molecular and Cellular Biology* **9**, 5022-5033.

Grady, D.L., Robert L. Ratliff, Donna LRobinson, Eric C.McCanlies, Julianne Meyne and Robert K. Moyzis (1992) *Proc. Natl. Acad. Sci. USA* **89**, 1695-1699.

Gross, D.S. and William T. Garrard (1986) *Molecular and Cellular Biology* **6**, 3010-3013.

Hamada, H., Marilanne G. Petrino and Takeo Kakunaga (1982) *Proc. Natl. Acad. Sci.* **79**, 6465-6469

Hamada, H., Michael Seidman, Bruce H. Howard and Cornelia M. Gorman (1984) *Molecular and Cellular Biology* **4**, 2622-2630.

Hamada, H. and Takeo Kakunaga (1982) *Nature* **298**, 396-398.

Hansen, R.S., Theresa K. Canfield, Mary M. Lamb, Stanley M. Gartler and Charles D. Laird (1993) *Cell* **73**, 1403-1409.

Hanzlik, A.J., MalgorzataM Osemlak-Hanzlik, Michael A.Hauser and David M.Kurnit (1993) *Nature Genetics* **3**, 44-48.

Herrera, V.L.M. and Nelson Ruiz-Opazo (1990) *The Journal of Biological Chemistry* **265,** 9555-9562.

Hirst, M.C., A. Barnicoat, G.Flynn, Q.Wang, M.Daker, V.J.Buckle, K.E.Davies and M.Bobrow (1993) *Human Molecular Genetics* **2,** 197-200.

Hornstra, I.K., David L Nelson, Stephen T. Warren and Thomas P. Yang (1993) *Human Molecular Genetics* **2**, 1659-1665.

Hutt, F.B. and W.F. Lamoreux (1940) *Heridity* **31**, 231-235.

Ionov, Y., Miguel A. Peinado, Sergei Malkhosyan, Darryl Shibata and Manuel Perucho (1993) *Nature* **363**, 558-561.

Jakowlew, S.B., Pamela J.Dillard, Michael B.Sporn and Anita B.Roberts (1988a) *Nucleic Acids Research* **16**, 8730.

Jakowlew, S.B., Pamela J. Dillard, Thomas S. Winokur, Kathleen C. Flanders, Michael B. Sporn and Anita B. Roberts (1990) *Developmental Biology* **143**, 1-14.

Jakowlew, S.B., Roert Lechleider, Andrew G.Geiser *et al.* (1992) *Molecular Endocrinology* **6**, 1-14.

Jakowlew, S.B., Pamela J. Dillard, Michael B. Sporn and Anita B. Roberts (1988b) *Molecular Endocrinology* **2**, 1186-1195.

Jakowlew, S.B., P.J.Dillard, P. Kondaiah, M.B. Sporn and A.B. Roberts (1988c) *Molecular Endocrinology* **2**, 747-755.

Jakowlew, S.B., Cubert, J., Danielpou, D., Sporn, M.B., Roberts, A.B. (1992) *Journal of Cellular Physiology* **150**, 377-385.

Jeffreys, J.D., Victoria Wilson and Swee Lay Thein (1985) *Nature* **314**, 67-73.

Jeffreys, A.J. (1985) *Nature* **314**, 67-73

Kandpal, R.P., Geeta Kandpal, Sherman M. Weissman (1994) *Proc. Natl. Acad. Sci. USA* **91**, 88-92.

Kao, C.C., Paul M. Lieberman, Martin C. Schmidt, Qiang Zhou, Rui Pei and Arnold J. Berk (1990) *Science* **248**, 1646-1650.

Karin, M., Alois Haslinger, Heidi Holtgreve, Robert I. Richards, Paul Krauter, Hannes M Westphal and Miguel Beato (1984) *Nature* **308**, 513-519.

Kladde, M.P., Yoshinori Kohwi, Terumi Kohwi-Shigematsu and Jack Gorski (1994) *Proc. Natl. Acad. Sci. USA* **91**, 1898-1902.

Klysik, J., Steven M. Stirdivant, Jackuelynn E. Larson, Phillip A. Hart and Robert D. Wells (1981) *Nature* **290**, 672-677.

Koide, R., T. Ikeuchi, O.Onodera, H. Tanaka *et al.* (1994) *Nature Genetics* **6**, 9-13.

Kondaiah, P., Ellen Van Obberghen-Schilling, Robert L.Ludwig, Ravi Dhart, Michael B. Sporn and Anita B.Roberts (1988) *The Journal of Biological Chemistry* **263**, 18313-18317.

Kondaiah, P., Mark J.Sands, Joseph M.Smith, Andy Fields, Anita B.Roberts, Michael B. Sporn and Douglas A.Melton (1990) *The Journal of Biological Chemistry* **265**, 1089-1093.

Kremer, B., P. Goldberg, S. E. Andrew, J. Theilmann, H. Telenius *et al.* (1994) *New England Journal of Medicine* **330**, 1401-1406.

Kristi, A. Wharton, Barry Yedvobnick, Victoria G. Finnerty and Spyros Artavanis-Tsakonas (1985) *Cell* **40,** 55-62.

Lancillotti, F., Manuel C. Lopez, Patricio Arias and Carlos Alonso (1987) *Proc. Natl. Acad. Sci. USA* **84**, 1560-1564.

Landerand, E.S. and David Botstein (1989) *Genetics* **121**, 185-199.

Lauri A. Aaltonen Päivi Peltomäki *et al.* (1993) *Science* **260**, 812-815.

Leach, F.S., Nicholas C. Nicolaides *et al.* (1993) *Cell* **75**, 1215-1225.

Levin, I., Lsa Santangelo, Hans Cheng, Lyman B. Crittenden and Jerry B. Dodgson (1994) *Journal of Heridity* **85**, 79-85.

Levinson, G. and George A. Gutman (1987) *Molecular Biological Evolution* **4**, 203-221.

Li, S.H., Melvin G. McInnis, Russell L. Margolis, Stylianos E. Antonarakis and Christopher A. Ross (1993) *Genomics* **16**, 572-579.

Lindblom, A., Ia Tannergard, Barbro Wereliu8s and Magnus Nordenskjöld (1993) *Nature Genetics* **5**, 279-282.

Lisitsyn, N., Natalya Lisitsyn, Michael Wigler (1993) *Science* **259**, 946-951.

Love, J.M., Andrew M. Knight, Marcia A. McAleer and John A. Todd (1990) *Nucleic Acids Research* **18**, 4123-4130.

Mahadevan, M., Catherine Tsilfidis, Luc Sabourin, Gary Shutler, Chris Amemiya *et al.* (1992) *Science* **255**, 1253-1255.

Martin, J.F., John J. Schwarz and Eric N. Olson (1993) *Proc. Natl. Acad. Sci. USA* **90**, 5282-5286.

Martin, G.B., Sergio H.Brommonschenkel, Julapark Chunwongse, Anne Frary, Martin W. Ganal, Rody Spivey, Tiyu Wu, Elizabeth D. Earle, Steven D. Tanksley (1993) *Science* 262, 1432-1436.

Massagué, J. (1990) *Annu. Rev. Cell Biology* **6,** 597-641.

Masatoshi Nei (1987) Molecular Evolutionary Genetics New York, Columbia University Press.

McDermott, J.C., M. Cristina Cardoso, Yie-Teh Yu, Vicente Andres *et al.* (1993) *Molecular and Cellular Biology* **13**, 2564-2577.

Michael Bishop, Steven M. Kappes, John W. Keele, Roger T. Stone, Sara L. F. Sunden, Gregory A. Hawkins, Sabina Solinas Toldo, Ruedi Fries, Michael D. Grosz, Jakyoung Yoo and Craig W. Beatie (1994) *Genetics* **136,** 619-639.

Mitchell, P.J. and Robert Tjian (1989) *Science* **245**, 371-378.

Moore, S.S., L.L.Sargenant, T.J. King, J.S. Mattick, M. Georges and D.J. Hetzel (1991) *Genomics* **10**, 654-660.

Moran, C. (1993) *Journal of Heridity* **84**, 274-280.

Moyzis, R.K., Judy M. Buckingham *et al.* (1988) *Proc. Natl. Acad. Sci. USA* **85**, 6622-6626.

Murphy, K.E. and James Stringer (1986) *Nucleic Acids Research* 7325-7340.

Nei, M. (1987) Molecular Evolutionary Genetics New York, Columbia University Press.

Nordheim, A., Mary Lou Pardue, Eileen M. Lafer, Achim Möller, B. David Stollar and Alexander Rich (1981) *Nature* **294**, 417-422

Nordheim, A. and Alexander Rich (1983a) *Nature* **303**, 674-679.

Nordheim, A. and Alexander Rich (1983b) *Proc. Natl. Acad. Sci. USA* **80**, 1821-1825.

O'Brien, S.J. (1993) *Current Opinion in Genetics and Development* **1**, 105-111.

Ondek, B., Lisa Gloss and Winship Herr (1988) *Nature* **333**, 40-45.

Ostrander, E.A., Pam M.Jong, Jasper Rine and Geoffrey Duyk (1992) *Proc. Natl. Acad. Sci. USA* **89**, 3419-3422.

Palazzolo, M.J., Stanley A. Sawyer, Christopher H. Martin, David A. Smoller and Daniel L. Hartl (1991) *Proc. Natl. Acad. Sci. USA* **88**, 8034-8038.

Pamela J. Mitchell and Robert Tjian (1989) *Science* **245**, 371-378.

Papadopoulas, N., Nicholas C. Nicolaides *et al.* (1994) *Science* **263**, 1625-1629.

Peck, L.J.(1985) *Cell* **49**, 129-137.

Petrukhin, K., S.G.Fischer *et al.* (1993) *Nature Genetics* **5**, 338343.

Pirrotta, V., E. Manet, E. Hardon, S. E. Bickel and M. Benson (1987) *The EMBL Journal* **6**, 791-799.

Pollock, R., and Richard Treisman (1991) *Genes and Development* **5**, 2327-2341.

Proufoot, N.J. and T. Maniatis (1980) *Cell* **21**, 537-544.

Ralph, W. (1991) Rural Research 150, 21-27.

Rassmann, K., Christian Schlötterer and Diethard Tautz (1981) *Electrophoresis* **12**, 113-118

Razin, A. and A. D. Riggs (1980) *Science* **210**, 604-610.

Regulski, M., Katherine Harding, Richard Kostriken, Francois Karch, Michael Levine and William McGinnis (1985) *Cell* **43,** 71-80.

Richards, R. I., K. Holman, S. Yu and G. R. Sutherland (1993) *Human Molecular Genetics* **2**, 1429-1435.

Riggins, G.J., Laurie K. Lokey, Jane L. Chastain, Harold A. Leiner, Stephanie L. Sherman, Keith D. Wilinson and Stephen T. Warren (1992) *Nature Genetics* **2**, 186-191.

Rik Derynck, Lucy Rhee, Ellson Y.Chen and Ann Van Tilburg (1987) *Nucleic Acids Research* **15**, 3188-3189.

Roberts, A.B., Seong-Jin Kim, Takafumi Noma *et al.* (1991) *Clinical Applications of TGFβ. Wiley, Chichester (Ciba Foundation Symposium 157)* p**7-28.**

Roger, J. (1983) *Nature* **305**, 101-102.

Rohrer, R.A., Leeson J. Alexander, John W. Keele, Tim P. Smith and Craig W. Beatie (1994) *Genetics* **136,** 231-245.

Sambrook, J., Fritsch, E.F, and Maniatis, T. (1989) Molecular Cloning: A Laboratory Manual. Cold Spring Harbor: Cold Spring Harbor Laboratory.

Santoro, C., Francesco Costanzo and Gennaro Ciliberto (1984) *The EMBL Journal* **3**, 1553-1559.

Seipel, K., Georgiev, O. and Schaffner, W. (1992) *EMBL J.* **11**, 4961-4968.

Service, R.F. (1994) *Science* **263**, 1559-1560.

Shibata, D., Miguel Angel Peinado, Yurij Ionov, Sergei Malkhosyan and Manuel Perucho (1994) *Nature Gemnetics* **6**, 273-281.

Siomi, H., Mikiko C. Siomi, Robert L. Nussbaum and Gideon Dreyfuss (1993) *Cell* **74**, 291-298.

Slightom, J.L., Ann E. Blechl and Oliver Smithies (1980) *Cell* **21**, 627-638.

Smith, G.P. (1976) *Science* **191**, 528-535.

Smith, G.R., Sam Martin Kunes, Dennis W. Schultz, Andrew Taylor and Kathleen L. Triman (1981) *Cell* **24**, 429-436.

Soller, M., Hans H. Cheng and Lyman B. Crittenden (Unpublished data).

Spada, A.R.L., Elizabeth M. Wilson, Dennis B. Lubahn, A. E. Harding and Kenneth H. Fischbeck (1991) *Nature* **352**, 77-79.

Spyros Artavanis-Tsakonas, Marc A. T. Muskavitch and Barry Yedvobnick (1983) *Proc. Natl. Acad. Sci. USA* **80**, 1977-1981.

Stallings, R.L., A.F. Ford, D. Nelson, D.C. Torney, C.E. Hiledbrand and R.K.Moyzis (1991) *Genetics* **10**, 807-815.

Stanton, L.W., Rosemary Watt and Kenneth B. Marcu (1983) *Nature* **303**, 401-406.

Steffen, P., A. Eggen, A. B. Dietz, J. E. Womack, G. Stranzinger, R. Fries (1993) *Animal Genetics* **24**, 121-124.

Streinmetz, M., Danielle Stephan and Kirsten Fischer Lindahl (1986) *Cell* **44**, 895-904.

Stringer, J.R. (1985) *Molecular and Cellular Biology* **5,** 1247-1259.

Sturtevant, A.H. (1913) *J. Exp. Zool.* **14**, 43-59.

Takeuchi K, Kawashima A, Nagafuchi A, Tsukita S (1994) *Journal of Cell Science* 107(7), 1921-1928.

Tanzi, R.E., K. Petrukhin, I. Chernov *et al.* (1993) *Nature Genetics* 5, 344-350.

Tautz, D. and M. Renz (1984) *Nucleic Acids Research* **12**, 4127-4138.

The Huntingto's Disease Collaborative Research Group (1993) *cell* **72,** 971-983.

Thibodeau, S.N., G. Bren, and D. Schaid (1993) *Science* **260**, 816-819.

Todd, J.A. *et al.* (1994) *Nature Genetics* 7, 390-395.

Treco, D.R., Barbara Thomas and Norman Arnheim (1985) *Molecular and Cellular Biology* **5**, 2029-2038

Treco, D.R. and Norman Arnheim (1986) *Molecular and Cellular Biology* **6**, 3934-3947.

Tuggle, C.K. *et al.* Future Genetics for the Animal Industry May, 1994-St.Louis Airport Hilton Hotel, St. Louis, MO.

Uberbacher, E.C. and R.J. Mural (1991) *Proc. Natl. Acad. Sci. USA* **88**, 11261-11265.

Valdes, A.M., Montgomery Slatkin and Nelson B. Freimer (1993) *Genetics* **133**, 737-749.

Verkerk, A.J., Maura Pieretti *et al.* (1991) *Cell* **65**, 905-914.

Wahls, W.P., Linda J. Wallace and Pter D. Moore (1990) *Molecular and Cellular Biology* **10**, 785-793.

Wakefield, L. *et al.* (1990) *J. Cell Science Supplement* **13**, 139-148.

Walmsly, R.M., Jack W. Szostak and Thomas D. Petes (1983) Nature 302, 84-86.

Wang Zhou and Donald D.Brown (1991) *Proc. Natl. Acad. Sci. USA* **88**, 11505-11509.

Wang, Y.H., Sorour Amirhaeri, Seongman Kang, Robert D. Wells and Jack D.Griffith (1994) *Science* **265**, 669-671.

Watson, J.D., Nancy H. Hopkins, Jeffrey W.Roberts, Joan Argetsinger Streitz and Alan M. Weiner (1987) *Molecular Biology of the Gene* Fourth Edition, J. Benjamin/ Cummings Publishing Company, Inc.

Weber, J.L. (1994) Nature Genetics 7, 343-350.

Weber, J.L. and Paul E. May (1989) *Am. J. Hum. Genet.* **44**, 388-396.

Weissenbach, J., Gabor Gyapay, Colette Dib, Alain Vignal, Jean Morissette, Philippe Millasseau, Guy Vaysseix and Mark Lathrop (1992) *Nature* **359**, 794-801.

Wells, R.G. and M. A. Hediger (1992) *Proc. Natl. Acad. Sci. USA* **89**, 5596-5600.

Wharton, K.A., Barry Yedvobnick, Victoria G. Finnerty and Spyros Artavanis-Tsakonas (1985) *Cell* **40,** 55-62.

Wilcox, A.S., Akbar S. Khan, Janet A. Hopkins, and James M. Sikela (1991) *Nucleic Acids Research* **19**, 1837-1843.

Wolfgang Stephan (1989) *Molecular Biological Evolution* **6**, 198-212.

Wolfgang Stephen and Soowon Cho (1994) *Genetics* **136**, 333-341.

Wong, A.K.C., H.A. Yee, J.H. van de Sande and J.B. Rattner (1990) *Chromosoma* **99**, 344-351

Yang, E.Y. and Harold L. Moses (1990) *The Journal of Cell Biology* **111**, 731-741.

Yu, Y.T., Roger E. Breitbart, Leslie B. Smoot, Youngsook Lee, Vijak Mahdavi and Bernardo Nadal-Ginard (1992) *Genes and Development* **6**, 1783-1798.